

Model Selection in High-Dimensional Misspecified Models *

Pallavi Basu¹, Yang Feng², and Jinchi Lv¹

¹University of Southern California

²Columbia University

Abstract

Model selection is indispensable to high-dimensional sparse modeling in selecting the best set of covariates among a sequence of candidate models. Most existing work assumes implicitly that the model is correctly specified or of fixed dimensions. Yet model misspecification and high dimensionality are common in real applications. In this paper, we investigate two classical Kullback-Leibler divergence and Bayesian principles of model selection in the setting of high-dimensional misspecified models. Asymptotic expansions of these principles reveal that the effect of model misspecification is crucial and should be taken into account, leading to the generalized AIC and generalized BIC in high dimensions. With a natural choice of prior probabilities, we suggest the generalized BIC with prior probability which involves a logarithmic factor of the dimensionality in penalizing model complexity. We further establish the consistency of the covariance contrast matrix estimator in a general setting. Our results and new method are supported by numerical studies.

Key Words: Model misspecification; high dimensionality; model selection; Kullback-Leibler divergence principle; Bayesian principle; AIC; BIC; GAIC; GBIC; GBIC_p.

1 Introduction

With rapid advances of modern technology, high-throughput data sets of unprecedented size, such as genetic and proteomic data, fMRI and functional data, and panel data in

*This work was partially supported by NSF CAREER Award DMS-0955316 and Grants DMS-0806030 and DMS-1308566.

economics and finance, are frequently encountered in many contemporary applications. In these applications, the dimensionality p can be comparable to or even much larger than the sample size n . A key assumption that often makes large-scale inference feasible is the sparsity of signals, meaning that only a small fraction of covariates contribute to the response when p is large compared to n . High-dimensional modeling with dimensionality reduction and feature selection plays an important role in these problems. A sparse modeling procedure typically produces a sequence of candidate models, each involving a possibly different subset of covariates. An important question is how to compare different models in high dimensions when models are possibly misspecified.

The problem of model selection has a long history with numerous contributions by many researchers. Among others, well-known model selection criteria are the AIC (Akaike, 1973 and 1974) and BIC (Schwarz, 1978), where the former is based on the Kullback-Leibler (KL) divergence principle of model selection and the latter is originated from the Bayesian principle. A great deal of work has been devoted to understanding and extending these methods. See, for example, Bozdogan (1987), Foster and George (1994), Konishi and Kitagawa (1996), Ing (2007), Chen and Chen (2008), Chen and Chan (2011), Ing and Lai (2011), Liu and Yang (2011), and Chang et al. (2014) in different model settings. The connections between the AIC and cross-validation have been investigated in Stone (1977), Hall (1990), and Peng et al. (2013) in various contexts. Model selection criteria such as AIC and BIC are frequently used for tuning parameter selection in regularization methods. For instance, model selection in the context of penalized likelihood methods has been studied in Fan and Li (2001), Wang et al. (2007), Wang et al. (2009), Zhang et al. (2010), and Fan and Tang (2013). In particular, Fan and Tang (2013) showed that classical information criteria such as AIC and BIC can be inconsistent for model selection when the dimensionality p grows very fast relative to sample size n .

Most existing work on model selection usually makes an implicit assumption that the model under study is correctly specified or of fixed dimensions. For example, White (1982) laid out a general theory of maximum likelihood estimation in misspecified models for the case of fixed dimensionality and independent and identically distributed (i.i.d.) observations. Recently, Lv and Liu (2014) investigated the problem of model selection with model misspecification and derived asymptotic expansions of both KL divergence and Bayesian principles

in misspecified generalized linear models, leading to the generalized AIC and generalized BIC, for the case of fixed dimensionality. A specific form of prior probabilities motivated by the KL divergence principle leads to the generalized BIC with prior probability (GBIC_p-L¹). Yet model misspecification and high dimensionality are both common in real applications. Thus a natural and important question is how to characterize the impact of model misspecification on model selection in high dimensions. We intend to provide some answer to this question in this paper. Our analysis enables us to suggest the generalized BIC with prior probability (GBIC_p) that involves a logarithmic factor of the dimensionality in penalizing model complexity.

To gain some insights into the challenges of the aforementioned problem, let us consider a motivating example. Assume that the response Y depends on the covariate vector $(X_1, \dots, X_p)^T$ through the functional form

$$Y = f(X_1) + f(X_2 - X_3) + f(X_4 - X_5) + \varepsilon, \quad (1)$$

where $f(x) = x^3/(x^2 + 1)$ and the remaining setting is as specified in Section 4.1.2. Consider sample size $n = 100$ and vary dimensionality p from 200 to 3200. Without prior knowledge about the true model structure, we take the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2)$$

as the working model, with the same notation therein, and apply some information criteria to hopefully recover the oracle working model consisting of the first five covariates. When $p = 200$, the traditional AIC and BIC, which ignore model misspecification, tend to select a model with size larger than five. As expected, GBIC_p-L works reasonably well by selecting the oracle working model half of the time. However, when p is increased to 3200, these methods fail to select such a model with significant probability and the prediction performance of the selected models deteriorates. This motivates us to study the problem of model selection in high-dimensional misspecified models. In contrast, our newly suggested GBIC_p can recover the oracle working model with significant probability in this challenging scenario.

The main contributions of our paper are threefold. First, we establish a systematic theory of model selection with model misspecification in high dimensions. The asymptotic

¹Here we use this notation to emphasize that the criterion is for the low-dimensional case, while reserving the original notation GBIC_p in Lv and Liu (2014) for the high-dimensional counterpart.

expansions for different model selection principles involve delicate and challenging technical analysis. Second, our work provides rigorous theoretical justification of the covariance contrast matrix estimator that incorporates the effect of model misspecification and is crucial for practical implementation. Such an estimator is shown to be consistent in the general setting of high-dimensional misspecified models. Third, we suggest the use of a new prior in the expansion for GBIC_p involving the $\log p$ term. This criterion has connections to the model selection criteria in Chen and Chen (2008) and Fan and Tang (2013) with the $\log p$ factor for the case of correctly specified models.

The rest of the paper is organized as follows. Section 2 introduces the setup for model misspecification. We present some key asymptotic properties of the quasi-maximum likelihood estimator and provide asymptotic expansions of KL divergence and Bayesian model selection principles in high dimensions in Section 3. Section 4 demonstrates the performance of different model selection criteria in high-dimensional misspecified models through several simulation and real data examples. We provide some discussions of our results and possible extensions in Section 5. The proofs of some main results are relegated to the Appendix. Additional technical proofs and numerical results are provided in the Supplementary Material.

2 Model misspecification

Assume that conditional on the covariates, the n -dimensional random response vector $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ has a true unknown distribution G_n with density function

$$g_n(\mathbf{y}) = \prod_{i=1}^n g_{n,i}(y_i), \quad (3)$$

where $\mathbf{y} = (y_1, \dots, y_n)^T$. Model (3) entails that all components of \mathbf{Y} are independent but not necessarily identically distributed. Consider a set of d covariates out of all p available covariates, where p can be much larger than n . Denote by \mathbf{X} the corresponding $n \times d$ deterministic design matrix. To simplify the technical presentation, we focus on the case of deterministic design. In practice, one chooses a family of working models to fit the data. Model misspecification generally occurs when the family of distributions is misspecified or some true covariates are missed.

Since the true model G_n is unknown, we choose a family of generalized linear models

(GLMs) $F_n(\cdot, \boldsymbol{\beta}) = F_n(\mathbf{z}; \mathbf{X}, \boldsymbol{\beta})$ with a canonical link as our working models, each of which has density function

$$f_n(\mathbf{z}, \boldsymbol{\beta}) d\mu_0(\mathbf{z}) = \prod_{i=1}^n f_0(z_i, \theta_i) d\mu_0(z_i) \equiv \prod_{i=1}^n \exp [z_i \theta_i - b(\theta_i)] d\mu(z_i), \quad (4)$$

where $\mathbf{z} = (z_1, \dots, z_n)^T$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)^T = \mathbf{X}\boldsymbol{\beta}$ with $\boldsymbol{\beta} \in \mathbb{R}^d$, $b(\theta)$ is a smooth convex function, μ_0 is the Lebesgue measure, and μ is some fixed measure on \mathbb{R} . Assume that $b''(\theta)$ is continuous and bounded away from 0, \mathbf{X} is of full column rank d , and EY_i^2 are bounded. Clearly $\{f_0(z, \theta) : \theta \in \mathbb{R}\}$ is a family of distributions in the regular exponential family and may not contain $g_{n,i}$'s.

To ease the presentation, define two vector-valued functions $\mathbf{b}(\boldsymbol{\theta}) = (b(\theta_1), \dots, b(\theta_n))^T$ and $\boldsymbol{\mu}(\boldsymbol{\theta}) = (b'(\theta_1), \dots, b'(\theta_n))^T$, and a matrix-valued function $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \text{diag}\{b''(\theta_1), \dots, b''(\theta_n)\}$. For any n -dimensional random vector \mathbf{Z} with distribution $F_n(\cdot, \boldsymbol{\beta})$ given by (4), it holds that $E\mathbf{Z} = \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta})$ and $\text{cov}(\mathbf{Z}) = \boldsymbol{\Sigma}(\mathbf{X}\boldsymbol{\beta})$. The density function (4) can be rewritten as

$$f_n(\mathbf{z}, \boldsymbol{\beta}) = \exp [\mathbf{z}^T \mathbf{X}\boldsymbol{\beta} - \mathbf{1}^T \mathbf{b}(\mathbf{X}\boldsymbol{\beta})] \prod_{i=1}^n \frac{d\mu}{d\mu_0}(z_i),$$

where $\frac{d\mu}{d\mu_0}$ denotes the Radon-Nikodym derivative. Given the observations \mathbf{y} and \mathbf{X} , this gives the quasi-log-likelihood function

$$\ell_n(\mathbf{y}, \boldsymbol{\beta}) = \log f_n(\mathbf{y}, \boldsymbol{\beta}) = \mathbf{y}^T \mathbf{X}\boldsymbol{\beta} - \mathbf{1}^T \mathbf{b}(\mathbf{X}\boldsymbol{\beta}) + \sum_{i=1}^n \log \frac{d\mu}{d\mu_0}(y_i). \quad (5)$$

The quasi-maximum likelihood estimator (QMLE) of the d -dimensional parameter vector $\boldsymbol{\beta}$ is defined as

$$\hat{\boldsymbol{\beta}}_n = \arg \max_{\boldsymbol{\beta} \in \mathbb{R}^d} \ell_n(\mathbf{y}, \boldsymbol{\beta}), \quad (6)$$

which is the solution to the score equation $\boldsymbol{\Psi}_n(\boldsymbol{\beta}) = \partial \ell_n(\mathbf{y}, \boldsymbol{\beta}) / \partial \boldsymbol{\beta} = \mathbf{X}^T [\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta})] = \mathbf{0}$. This equation becomes the normal equation $\mathbf{X}^T \mathbf{y} = \mathbf{X}^T \mathbf{X}\boldsymbol{\beta}$ in the linear regression model.

The KL divergence (Kullback and Leibler, 1951) of the model $F_n(\cdot, \boldsymbol{\beta})$ from the true model G_n can be written as $I(g_n; f_n(\cdot, \boldsymbol{\beta})) = E \log g_n(\mathbf{Y}) - E \ell_n(\mathbf{Y}, \boldsymbol{\beta})$. The best working model that is closest to the true model under the KL divergence has parameter vector $\boldsymbol{\beta}_{n,0} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^d} I(g_n; f_n(\cdot, \boldsymbol{\beta}))$, which solves the equation

$$\mathbf{X}^T [E\mathbf{Y} - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta})] = \mathbf{0}. \quad (7)$$

We introduce two matrices that play a key role in model selection with model misspecification. Define

$$\text{cov} [\Psi_n(\beta_{n,0})] = \text{cov} (\mathbf{X}^T \mathbf{Y}) = \mathbf{X}^T \text{cov}(\mathbf{Y}) \mathbf{X} = \mathbf{B}_n \quad (8)$$

with $\text{cov}(\mathbf{Y}) = \text{diag}\{\text{var}(Y_1), \dots, \text{var}(Y_n)\}$ by the independence assumption,

$$\frac{\partial^2 I(g_n; f_n(\cdot, \beta))}{\partial \beta^2} = -\frac{\partial^2 \ell_n(\mathbf{y}, \beta)}{\partial \beta^2} = \mathbf{X}^T \Sigma(\mathbf{X}\beta) \mathbf{X} = \mathbf{A}_n(\beta), \quad (9)$$

and $\mathbf{A}_n = \mathbf{A}_n(\beta_{n,0})$. Observe that \mathbf{A}_n and \mathbf{B}_n are the covariance matrices of $\mathbf{X}^T \mathbf{Y}$ under the best misspecified GLM $F_n(\cdot, \beta_{n,0})$ and the true model G_n , respectively.

3 High-dimensional model selection in misspecified models

We now present the asymptotic expansions of both KL divergence and Bayesian model selection principles in high-dimensional misspecified GLMs.

3.1 Technical conditions and asymptotic properties of QMLE in high dimensions

We list a few technical conditions required to prove the asymptotic properties of QMLE with diverging dimensionality. Denote by $\|\cdot\|_2$ the vector L_2 -norm and the matrix operator norm.

Condition 1. *There exists some constant $H > 0$ such that for each $1 \leq i \leq n$, $P(|q_i| > t) \leq H \exp(-t^2/H)$ for any $t \geq 0$, where $(q_1, \dots, q_n)^T = \text{cov}(\mathbf{Y})^{-1/2}(\mathbf{Y} - E\mathbf{Y})$.*

Condition 2. *There exist positive constants $c_1, c_0 > 8c_1^{-2}H$, and $r < 1/4$ such that for sufficiently large n , $\min_{\beta \in N_n(\delta_n)} \lambda_{\min} \{\mathbf{V}_n(\beta)\} > c_1 n^{-r}$ and $\lambda_{\min}(\mathbf{B}_n) \gg d\delta_n^2$, where $\delta_n = n^r(c_0 \log n)^{1/2}$, $N_n(\delta_n) = \{\beta \in \mathbb{R}^d : \|(n^{-1}\mathbf{B}_n)^{1/2}(\beta - \beta_{n,0})\|_2 \leq (n/d)^{-1/2}\delta_n\}$, and $\mathbf{V}_n(\beta) = \mathbf{B}_n^{-1/2} \mathbf{A}_n(\beta) \mathbf{B}_n^{-1/2}$. Moreover, $d = o\{n^{(1-4r)/3}(\log n)^{-2/3}\}$.*

Condition 3. *Assume $\sum_{i=1}^n (\mathbf{x}_i^T \mathbf{B}_n^{-1} \mathbf{x}_i)^{3/2} = o(1)$ and $\max_{1 \leq i \leq n} E|Y_i - EY_i|^3 = O(1)$.*

Condition 4. *Assume*

$$\max_{\beta_1, \dots, \beta_d \in N_n(\delta_n)} \|\tilde{\mathbf{V}}_n(\beta_1, \dots, \beta_d) - \mathbf{V}_n\|_2 = O(dn^{-1/2}\delta_n),$$

where $\mathbf{V}_n = \mathbf{V}_n(\boldsymbol{\beta}_{n,0}) = \mathbf{B}_n^{-1/2} \mathbf{A}_n \mathbf{B}_n^{-1/2}$ and $\tilde{\mathbf{V}}_n(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_d) = \mathbf{B}_n^{-1/2} \tilde{\mathbf{A}}_n(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_d) \mathbf{B}_n^{-1/2}$ with $\tilde{\mathbf{A}}_n(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_d)$ a $d \times d$ matrix with j th row the corresponding row of $\mathbf{A}_n(\boldsymbol{\beta}_j)$ for each $1 \leq j \leq d$. Moreover, $\lambda_{\max}(\mathbf{V}_n)$ is a polynomial order of n .

Conditions 1 and 2 are some basic assumptions for establishing the consistency of the QMLE $\hat{\boldsymbol{\beta}}_n$ in Theorem 1. In particular, Condition 1 assumes that the standardized response has sub-Gaussian distribution which facilitates the derivation of the deviation probability bound. Conditions 2–4 are similar to those in Lv and Liu (2014), except for some major differences due to the high-dimensional setting. In particular, Condition 2 allows the minimum eigenvalue of $\mathbf{V}_n(\boldsymbol{\beta})$ to converge to zero at a certain rate as n increases in a neighborhood $N_n(\delta_n)$ of $\boldsymbol{\beta}_{n,0}$. Such a neighborhood is wider compared to that for the case of fixed dimensionality. The dimensionality d of the QMLE is allowed to diverge with n . Conditions 3 and 4 are imposed to establish the asymptotic normality of $\hat{\boldsymbol{\beta}}_n$.

Theorem 1. (Consistency of QMLE). *Under Conditions 1–2, the QMLE $\hat{\boldsymbol{\beta}}_n$ satisfies $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0} = o_P(1)$ and further $\hat{\boldsymbol{\beta}}_n \in N_n(\delta_n)$ with probability $1 - O(n^{-\alpha})$ for some large positive constant α .*

Theorem 2. (Asymptotic normality). *Under Conditions 1–4, the QMLE $\hat{\boldsymbol{\beta}}_n$ satisfies*

$$\mathbf{D}_n \mathbf{C}_n (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, I_m),$$

where $\mathbf{C}_n = \mathbf{B}_n^{-1/2} \mathbf{A}_n$ and \mathbf{D}_n is any $m \times d$ matrix such that $\mathbf{D}_n \mathbf{D}_n^T = I_m$.

Theorems 1 and 2 establish the consistency and asymptotic normality of the QMLE in high-dimensional misspecified GLM. These results provide the theoretical foundation for the technical analyses in Sections 3.2–3.4. The asymptotic theory of the QMLE reduces to that of the maximum likelihood estimator (MLE) when the model is correctly specified. Our results extend those in Lv and Liu (2014) for the case of fixed dimensionality. We next introduce a few additional conditions for deriving the asymptotic expansions of the two model selection principles.

Condition 5. *There exists some constant α_1 with $0 < \alpha_1 < \alpha/2 - 1$ such that $b''(\cdot) = O(n^{\alpha_1})$ and for sufficiently large n , $N_n(\delta_n) \subset M_n(\alpha_1) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|\mathbf{X}\boldsymbol{\beta}\|_\infty \leq \alpha_1 \log n\}$, where constant α is given in Theorem 1.*

Condition 6. Assume that $\pi(h(\boldsymbol{\beta})) = \frac{d\mu_{\mathfrak{M}}}{d\mu_0}(h(\boldsymbol{\beta}))$ satisfies

$$\inf_{\boldsymbol{\beta} \in N_n(2\delta_n)} \pi(h(\boldsymbol{\beta})) \geq c_2 \text{ and } \sup_{\boldsymbol{\beta} \in \mathbb{R}^d} \pi(h(\boldsymbol{\beta})) \leq c_3 \quad (10)$$

with $c_2, c_3 > 0$ some constants, and $\rho_n(\delta_n) = \max_{\boldsymbol{\beta} \in N_n(2\delta_n)} \max\{|\lambda_{\min}(\mathbf{V}_n(\boldsymbol{\beta}) - \mathbf{V}_n)|, |\lambda_{\max}(\mathbf{V}_n(\boldsymbol{\beta}) - \mathbf{V}_n)|\} = o\{n^{-(1-r)/3}\}$.

Condition 7. Assume that $n^{-1}\mathbf{A}_n(\boldsymbol{\beta})$, $n^{-1}\mathbf{X}^T \text{diag}\{|\boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}) - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})|\}\mathbf{X}$, and $n^{-1}\mathbf{X}^T \text{diag}\{[\boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}) - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})] \circ [\boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}) - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})]\}\mathbf{X}$ are Lipschitz (in operator norm) with constant $L > 0$ in $N_n(\delta_n)$, and $\|\mathbf{X}\|_{\infty} = O(n^{\alpha_2})$ with constant $0 \leq \alpha_2 < r$, where \circ represents the Hadamard (componentwise) product and $\|\cdot\|_{\infty}$ denotes the entrywise matrix L_{∞} -norm.

Condition 8. Assume $\sum_{i=1}^n \{[EY_i - (\boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0}))_i]^2 / \text{var}(Y_i)\}^2 = O(n^{\alpha_3})$ with some constant $0 \leq \alpha_3 \leq 4(r - \alpha_2)$.

The first part of Condition 5 holds naturally for linear and logistic regression models, and is introduced to accommodate the case of Poisson regression. The second part of Condition 5 is a mild assumption ensuring that the restricted QMLE coincides with its unrestricted version with significant probability, which is key to the asymptotic expansion of the KL divergence principle in high dimensions in Theorem 3. It is worth mentioning that the set $M_n(\alpha_1)$ grows with n , while the neighborhood $N_n(\delta_n)$ is asymptotically shrinking.

Condition 6 is similar to the one in Lv and Liu (2014), except that we need to specify the rate at which $\rho_n(\delta_n)$ converges to zero. Condition 7 requires the Lipschitz property for those matrix-valued functions. The bound on the entry-wise matrix L_{∞} -norm of the design matrix is mild. Condition 8 is a sensible assumption bounding the effect of the model bias. In particular, Conditions 7 and 8 are introduced only for proving the consistency of the covariance contrast matrix in the general setting in Theorem 4.

3.2 Generalized AIC in misspecified models

Given a sequence of subsets $\{\mathfrak{M}_m : m = 1, \dots, M\}$ of the full model $\{1, \dots, p\}$, we can construct a sequence of QMLE's $\{\hat{\boldsymbol{\beta}}_{n,m} : m = 1, \dots, M\}$ by fitting the GLM (4). A natural question is how to compare those fitted models. The QMLEs $\{\hat{\boldsymbol{\beta}}_{n,m} : m = 1, \dots, M\}$ become the MLEs when the model is correctly specified.

Akaike's principle of model selection is choosing the model \mathfrak{M}_{m_0} that minimizes the KL divergence $I(g_n; f_n(\cdot, \hat{\beta}_{n,m}))$ of the fitted model $F_n(\cdot, \hat{\beta}_{n,m})$ from the true model G_n , that is,

$$m_0 = \arg \min_{m \in \{1, \dots, M\}} I(g_n; f_n(\cdot, \hat{\beta}_{n,m})), \quad (11)$$

where

$$I(g_n; f_n(\cdot, \hat{\beta}_{n,m})) = E \log g_n(\tilde{\mathbf{Y}}) - \eta_n(\hat{\beta}_{n,m}) \quad (12)$$

with $\eta_n(\beta) = E \ell_n(\tilde{\mathbf{Y}}, \beta)$ and $\tilde{\mathbf{Y}}$ an independent copy of \mathbf{Y} . Thus

$$m_0 = \arg \max_{m \in \{1, \dots, M\}} \eta_n(\hat{\beta}_{n,m}) = \arg \max_{m \in \{1, \dots, M\}} E_{\tilde{\mathbf{Y}}} \ell_n(\tilde{\mathbf{Y}}, \hat{\beta}_{n,m}),$$

which shows that Akaike's principle of model selection is equivalent to choosing the model \mathfrak{M}_{m_0} that maximizes the expected log-likelihood with the expectation taken with respect to an independent copy of \mathbf{Y} . Using the asymptotic theory of MLE, Akaike (1973) showed that for the case of i.i.d. observations, $\eta_n(\hat{\beta}_n)$ can be asymptotically expanded as $\ell_n(\mathbf{y}, \hat{\beta}_n) - |\mathfrak{M}|$, which leads to the seminal AIC for comparing competing models:

$$\text{AIC}(\mathbf{y}, \mathfrak{M}) = -2\ell_n(\mathbf{y}, \hat{\beta}_n) + 2|\mathfrak{M}|. \quad (13)$$

For simplicity, we drop the last term in (5) which does not depend on β , and redefine the quasi-log-likelihood as $\ell_n(\mathbf{y}, \beta) = \mathbf{y}^T \mathbf{X}\beta - \mathbf{1}^T \mathbf{b}(\mathbf{X}\beta)$ hereafter.

Theorem 3. *Under Conditions 1–5, we have with probability tending to one,*

$$E\eta_n(\hat{\beta}_n) = E\ell_n(\mathbf{y}, \hat{\beta}_n) - \text{tr}(\mathbf{H}_n) + o\{\text{tr}(\mathbf{H}_n)\}, \quad (14)$$

where $\mathbf{H}_n = \mathbf{A}_n^{-1} \mathbf{B}_n$.

Theorem 3 generalizes the corresponding result in Lv and Liu (2014) to high dimensions. However, we would like to point out that our new technical analysis differs substantially from theirs due to the challenges of diverging dimensionality. The asymptotic expansion in Theorem 3 enables us to introduce the generalized AIC (GAIC) as follows.

Definition 1. *We define GAIC of model \mathfrak{M} as*

$$\text{GAIC}(\mathbf{y}, \mathfrak{M}; F_n) = -2\ell_n(\mathbf{y}, \hat{\beta}_n) + 2\text{tr}(\hat{\mathbf{H}}_n), \quad (15)$$

where $\hat{\mathbf{H}}_n$ is a consistent estimator of \mathbf{H}_n specified in Section 3.3.

When the model is correctly specified, it holds that $\text{tr}(\widehat{\mathbf{H}}_n) \approx \text{tr}(\mathbf{I}_d) = |\mathfrak{M}|$, under which GAIC reduces to AIC asymptotically. We demonstrate in the simulation studies that GAIC can improve over the original AIC substantially in the presence of model misspecification.

3.3 Estimation of covariance contrast matrix

From the asymptotic expansions for the GAIC, GBIC, and GBIC_p (the latter two to be introduced in Section 3.4), a common term is the covariance contrast matrix \mathbf{H}_n , which characterizes the impact of model misspecification. Therefore, providing an accurate estimator for such a matrix \mathbf{H}_n is of vital importance in the application of these information criteria.

Consider the plug-in estimator $\widehat{\mathbf{H}}_n = \widehat{\mathbf{A}}_n^{-1} \widehat{\mathbf{B}}_n$ with $\widehat{\mathbf{A}}_n$ and $\widehat{\mathbf{B}}_n$ defined as follows. Since the QMLE $\widehat{\boldsymbol{\beta}}_n$ provides a consistent estimator of $\boldsymbol{\beta}_{n,0}$ in the best misspecified GLM $F_n(\cdot, \boldsymbol{\beta}_{n,0})$, a natural estimate of matrix \mathbf{A}_n is given by

$$\widehat{\mathbf{A}}_n = \mathbf{A}_n(\widehat{\boldsymbol{\beta}}_n) = \mathbf{X}^T \boldsymbol{\Sigma}(\mathbf{X} \widehat{\boldsymbol{\beta}}_n) \mathbf{X}. \quad (16)$$

When the model is correctly specified, the following simple estimator

$$\widehat{\mathbf{B}}_n = \mathbf{X}^T \text{diag} \left\{ \left[\mathbf{y} - \boldsymbol{\mu}(\mathbf{X} \widehat{\boldsymbol{\beta}}_n) \right] \circ \left[\mathbf{y} - \boldsymbol{\mu}(\mathbf{X} \widehat{\boldsymbol{\beta}}_n) \right] \right\} \mathbf{X} \quad (17)$$

gives an asymptotically unbiased estimator of \mathbf{B}_n .

Theorem 4. (*Consistency of estimator*) Assume that Conditions 1–3 and 7–8 hold, the eigenvalues of $n^{-1} \mathbf{A}_n$ and $n^{-1} \mathbf{B}_n$ are bounded away from 0 and ∞ , and $d = o\{n^{(1-4r)/4}\}$. Then the plug-in estimator $\widehat{\mathbf{H}}_n$ satisfies $\text{tr}(\widehat{\mathbf{H}}_n) = \text{tr}(\mathbf{H}_n) + o_P(1)$ and $\log |\widehat{\mathbf{H}}_n| = \log |\mathbf{H}_n| + o_P(1)$.

Theorem 4 improves the result in Lv and Liu (2014) in two important aspects. First, the consistency of the covariance contrast matrix estimator was previously justified in Lv and Liu (2014) for the case of correctly specified model. Our new result shows that the simple plug-in estimator $\widehat{\mathbf{H}}_n$ still enjoys consistency in the general setting of model misspecification. Second, the result in Theorem 4 holds for the case of diverging dimensionality. These theoretical guarantees are crucial to the practical implementation of those information criteria. Our numerical studies reveal that such an estimate works well in a variety of model misspecification settings.

3.4 Generalized BIC in misspecified models

Given a set of competing models $\{\mathfrak{M}_m : m = 1, \dots, M\}$, a popular Bayesian model selection procedure is to first put nonzero prior probability $\alpha_{\mathfrak{M}_m}$ on each model \mathfrak{M}_m , and then choose a prior distribution $\mu_{\mathfrak{M}_m}$ for the parameter vector in the corresponding model. Assume that the density function of $\mu_{\mathfrak{M}_m}$ is bounded in $\mathbb{R}^{\mathfrak{M}_m} = \mathbb{R}^{d_m}$ with $d_m = |\mathfrak{M}_m|$ and locally bounded away from zero throughout the domain. The Bayesian principle of model selection is to choose the most probable model *a posteriori*, that is, choose model \mathfrak{M}_{m_0} such that

$$m_0 = \arg \max_{m \in \{1, \dots, M\}} S(\mathbf{y}, \mathfrak{M}_m; F_n), \quad (18)$$

where the log-marginal-likelihood is

$$S(\mathbf{y}, \mathfrak{M}_m; F_n) = \log \int \alpha_{\mathfrak{M}_m} \exp[\ell_n(\mathbf{y}, \boldsymbol{\beta})] d\mu_{\mathfrak{M}_m}(\boldsymbol{\beta}) \quad (19)$$

with the log-likelihood $\ell_n(\mathbf{y}, \boldsymbol{\beta})$ as in (5) and the integral over \mathbb{R}^{d_m} .

To ease the presentation, for any $\boldsymbol{\beta} \in \mathbb{R}^d$ we define a quantity

$$\ell_n^*(\mathbf{y}, \boldsymbol{\beta}) = \ell_n(\mathbf{y}, \boldsymbol{\beta}) - \ell_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n), \quad (20)$$

which is the deviation of the quasi-log-likelihood from its maximum. Then from (19) and (20), we have

$$S(\mathbf{y}, \mathfrak{M}_m; F_n) = \ell_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) + \log E_{\mu_{\mathfrak{M}_m}}[U_n(\boldsymbol{\beta})^n] + \log \alpha_{\mathfrak{M}_m}, \quad (21)$$

where $U_n(\boldsymbol{\beta}) = \exp[n^{-1}\ell_n^*(\mathbf{y}, \boldsymbol{\beta})]$.

Theorem 5. *Under Conditions 1–3 and 6, we have with probability tending to one,*

$$\begin{aligned} S(\mathbf{y}, \mathfrak{M}; F_n) &= \ell_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) - \frac{\log n}{2} |\mathfrak{M}| + \frac{1}{2} \log |\mathbf{H}_n| + \log \alpha_{\mathfrak{M}} \\ &\quad + \frac{\log(2\pi)}{2} |\mathfrak{M}| + \log c_n + o(1), \end{aligned} \quad (22)$$

where $\mathbf{H}_n = \mathbf{A}_n^{-1} \mathbf{B}_n$ and $c_n \in [c_2, c_3]$.

The asymptotic expansion of the Bayes factor in Theorem 5 leads us to introduce the generalized BIC (GBIC) as follows.

Definition 2. We define GBIC of model \mathfrak{M} as

$$GBIC(\mathbf{y}, \mathfrak{M}; F_n) = -2\ell_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) + (\log n)|\mathfrak{M}| - \log |\hat{\mathbf{H}}_n|, \quad (23)$$

where $\hat{\mathbf{H}}_n$ is a consistent estimator of \mathbf{H}_n .

It is clear from (23) that GBIC contains an extra term compared to BIC that replaces the factor 2 with $\log n$ in penalizing model complexity in (13). This additional term reflects the effect of model misspecification. When the model is correctly specified, GBIC reduces to BIC asymptotically.

The choice of the prior probabilities $\alpha_{\mathfrak{M}_m}$ is important in high dimensions. Lv and Liu (2014) suggested prior probability $\alpha_{\mathfrak{M}_m} \propto e^{-D_m}$ for each candidate model \mathfrak{M}_m , where the quantity D_m is defined as

$$D_m = E \left[I(g_n; f_n(\cdot, \hat{\boldsymbol{\beta}}_{n,m})) - I(g_n; f_n(\cdot, \boldsymbol{\beta}_{n,m,0})) \right] \quad (24)$$

and the subscript m indicates a particular candidate model. The motivation is that the further the QMLE $\hat{\boldsymbol{\beta}}_{n,m}$ is away from the best misspecified GLM $F_n(\cdot, \boldsymbol{\beta}_{n,m,0})$, the lower prior we assign to that model. In the high-dimensional setting when p can be much larger than n , it is sensible to take into account the complexity of the space of all possible sparse models with the same size as \mathfrak{M}_m . This observation motivates us to consider a new prior of the form

$$\alpha_{\mathfrak{M}_m} \propto \left(\frac{p}{d} \right)^{-1} e^{-D_m} \quad (25)$$

with $d = |\mathfrak{M}_m|$. Such a complexity factor has been exploited in the extended BIC (EBIC) in Chen and Chen (2008), who showed that using the term $\left(\frac{p}{d} \right)^{-\gamma}$ with some constant $0 < \gamma \leq 1$, the EBIC can be model selection consistent for $p = O(n^\kappa)$ with some positive constant κ satisfying $1 - (2\kappa)^{-1} < \gamma$.

Under the assumption of $d = o(p)$, an application of Stirling's formula shows that up to an additive constant, it holds that $\log \alpha_{\mathfrak{M}_m} \approx -D_m - d \log p - d + d \log d$. Thus for the prior defined in (25), we have an additional term $-(\log p + 1 - \log d)|\mathfrak{M}|$ in the asymptotic expansion for GBIC. When p is of order n^κ with some constant $\kappa > 0$, this new term is of the same order as $-(\log n)|\mathfrak{M}|$. When $\log p$ is of order n^κ with some constant $0 < \kappa < 1$, the $\log p$ term dominates that involving $\log n$. Fan and Tang (2013) proposed a similar term

$\log(\log n) \log p$ term to ameliorate the BIC for the case of correctly specified models with non-polynomially growing dimensionality p . The following theorem provides the asymptotic expansion of the Bayes factor with the particular choice of prior in (25).

Theorem 6. *Assume that Conditions 1–6 hold, $\alpha_{\mathfrak{M}_m} = C \binom{p}{d}^{-1} e^{-D_m}$ with $C > 0$ some normalization constant, and $d = o(p)$. Then we have with probability tending to one,*

$$\begin{aligned} S(\mathbf{y}, \mathfrak{M}; F_n) &= \ell_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) - (\log p^*)|\mathfrak{M}| - \frac{1}{2} \text{tr}(\mathbf{H}_n) + \frac{1}{2} \log |\mathbf{H}_n| \\ &\quad + \log(Cc_n) + o(1), \end{aligned} \quad (26)$$

where $\mathbf{H}_n = \mathbf{A}_n^{-1} \mathbf{B}_n$, $p^* = \max\{n, p\}$, and $c_n \in [c_2, c_3]$.

Similarly to the GBIC, we now define a new information criterion, the generalized BIC with prior probability (GBIC_p), based on Theorem 6.

Definition 3. *We define GBIC_p of model \mathfrak{M} as*

$$\text{GBIC}_p(\mathbf{y}, \mathfrak{M}; F_n) = -2\ell_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) + 2(\log p^*)|\mathfrak{M}| + \text{tr}(\hat{\mathbf{H}}_n) - \log |\hat{\mathbf{H}}_n|, \quad (27)$$

where $\hat{\mathbf{H}}_n$ is a consistent estimator of \mathbf{H}_n .

In correctly specified models, the term $\text{tr}(\hat{\mathbf{H}}_n) - \log |\hat{\mathbf{H}}_n|$ is asymptotically close to $|\mathfrak{M}|$ when $\hat{\mathbf{H}}_n$ is a consistent estimator of $\mathbf{H}_n = \mathbf{I}_d$. Thus compared to BIC with factor $\log n$, the GBIC_p contains a larger factor $\log p$ when p grows non-polynomially with n . This leads to a heavier penalty on model complexity similarly as in Fan and Tang (2013). As pointed out in Lv and Liu (2014), the right hand side of (27) can be viewed as a sum of three terms: the goodness of fit, model complexity, and model misspecification. An important distinction with the low-dimensional counterpart of GBIC_p is that our new criterion explicitly takes into account the dimensionality of the whole feature space.

4 Numerical studies

The asymptotic expansions of both KL divergence and Bayesian principles in Section 3 have enabled us to introduce the GAIC, GBIC, and GBIC_p for model selection in high dimensions with model misspecification. We now investigate their performance in comparison to the

information criteria AIC, BIC, and $\text{GBIC}_p\text{-L}$ in high-dimensional misspecified models via simulation examples as well as two real data sets. For each simulation study, we set the number of repetitions to be 100 and examined the scenarios when the dimensionality grows ($p = 200, 400, 1600$, and 3200).

4.1 Simulation examples

4.1.1 Sparse linear regression with interaction and weak effects

The first model we consider is the following high-dimensional linear regression model with interaction and weak effects

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{x}_{p+1} + \boldsymbol{\varepsilon}, \quad (28)$$

where $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p)$ is an $n \times p$ design matrix, $\mathbf{x}_{p+1} = \mathbf{x}_1 \circ \mathbf{x}_2$ is an interaction term which is the product of the first two covariates, the rows of \mathbf{X} are sampled as i.i.d. copies from $N(\mathbf{0}, I_p)$, and the error vector $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$. We set $\boldsymbol{\beta}_0 = (1, -1.25, 0.75, -0.95, 1.5, 0.1, -0.1, 0.1, -0.1, 0.1, 0, \dots, 0)^T$, $n = 100$, and $\sigma = 0.25$. Although the data was generated from model (28), we fit the linear regression model (2) without interaction, which is a typical example of model misspecification. In view of (28), the true model involves only the first ten covariates in a nonlinear form. Since the other covariates are independent of those ten covariates, the oracle working model is $\text{supp}(\boldsymbol{\beta}_0) = \{1, \dots, 10\}$ as argued in Lv and Liu (2014). Due to the high dimensionality, it is computationally prohibitive to implement the best subset selection. Therefore, we first applied the regularization method SICA (Lv and Fan, 2009) to build a sequence of sparse models and then selected the final model using a model selection criterion. In practice, one can apply any preferred variable selection procedure to obtain a sequence of candidate models.

In addition to comparing the models selected by different information criteria, we also considered the estimate based on the oracle working model $M_0 = \{1, \dots, 10\}$ as a benchmark and used both measures of prediction and variable selection. Denote by \widehat{M} the selected model. We split the oracle working model into the set of strong effects $M_{0,s} = \{1, \dots, 5\}$ and that of weak effects $M_{0,w} = \{6, \dots, 10\}$. It is interesting to observe that all criteria tend to miss the entire set of weak effects $M_{0,w}$ due to their very low signal strength. Therefore, we focused on comparing the model selection performance in recovering the set of strong effects $M_{0,s}$.

We report the strong effect consistent selection probability (the portion of simulations where $\widehat{M} = M_{0,s}$), the strong effect inclusion probability (the portion of simulations where $\widehat{M} \supset M_{0,s}$), and the prediction error $E(Y - \mathbf{x}^T \widehat{\beta})^2$ with $\widehat{\beta}$ an estimate and (\mathbf{x}^T, Y) an independent observation. To evaluate the prediction performance of different criteria, we calculated the average prediction error on an independent test sample of size 10,000. The results for prediction error and model selection performance are summarized in Table 1. To save space, the number of false positives $|\widehat{M} \cap M_0^c|$ and the numbers of false negatives for strong effects $|\widehat{M}^c \cap M_{0,s}|$ and weak effects $|\widehat{M}^c \cap M_{0,w}|$, respectively, are reported in Table 6 in the Supplementary Material.

It is clear that as the dimensionality p increases, the consistent selection probability tends to decrease and the prediction error tends to increase for all information criteria. Generally speaking, GAIC improved over AIC, and GBIC, GBIC_p-L, and GBIC_p performed better than BIC in terms of both prediction and variable selection. In particular, the model selected by our new information criterion GBIC_p delivered the best performance with the smallest prediction error and highest strong effect consistent selection probability across all settings.

Meanwhile it is also interesting to see what results different model selection criteria lead to when the model is correctly specified. To this end, we regenerate the solution path based on the linear regression model with the interaction $\mathbf{x}_{p+1} = \mathbf{x}_1 \circ \mathbf{x}_2$ added. The same performance measures are calculated for this scenario with the results reported in Tables 2 and 7, where the latter table is included in the Supplementary Material. A comparison of these results with those in Tables 1 and 6 gives several interesting observations. First, all model selection criteria have a better performance when the model is correctly specified in terms of both model selection and prediction. Second, it is worth noting that while all model selection criteria except AIC work reasonably well for the correctly specified model, all but the newly proposed GBIC_p have a very low consistent selection probability under both model misspecification and high dimensionality. Third, it is interesting to see that GBIC_p outperforms the existing methods even under the correctly specified model in terms of consistent selection probability.

Table 1: Simulation results for Example 4.1.1 with all entries multiplied by 100 when the model is misspecified, with the oracle results based on both strong effects and weak effects.

Strong effect consistent selection probability with inclusion probability							
p	AIC	BIC	GAIC	GBIC	GBIC _{p} -L	GBIC _{p}	Oracle
200	0(99)	29(99)	21(99)	32(99)	67(98)	73(98)	100(100)
400	0(100)	9(100)	8(100)	19(100)	54(100)	76(100)	100(100)
1600	0(100)	0(100)	9(100)	0(100)	27(100)	66(100)	100(100)
3200	0(100)	0(100)	4(100)	0(100)	16(100)	64(100)	100(100)
Median prediction error with robust standard deviation in parentheses							
200	164(35)	130(13)	130(10)	128(12)	125(8)	125(8)	121(7)
400	162(29)	154(38)	129(13)	131(22)	125(9)	122(10)	120(7)
1600	168(31)	172(28)	134(13)	170(28)	129(14)	125(10)	121(7)
3200	159(22)	169(23)	135(14)	167(23)	134(15)	125(13)	120(8)

Table 2: Simulation results for Example 4.1.1 with all entries multiplied by 100 when the model is correctly specified, with the oracle results based on both strong effects and weak effects.

Strong effect consistent selection probability with inclusion probability							
p	AIC	BIC	GAIC	GBIC	GBIC _{p} -L	GBIC _{p}	Oracle
200	2(100)	82(100)	81(99)	82(100)	87(100)	91(100)	100(100)
400	8(100)	76(100)	76(100)	84(100)	90(100)	94(100)	100(100)
1600	39(95)	74(99)	65(89)	79(99)	88(100)	96(100)	100(100)
3200	64(94)	84(98)	72(88)	84(98)	94(100)	95(100)	100(100)
Median prediction error with robust standard deviation (RSD) in parentheses							
200	13.6(1.9)	11.2(1.0)	11.2(1.0)	11.2(1.0)	11.6(1.3)	11.7(1.2)	7.0(0.4)
400	12.1(1.4)	11.5(1.3)	11.5(1.2)	11.5(1.2)	11.7(1.3)	11.8(1.0)	6.9(0.4)
1600	12.4(8.3)	12.0(7.9)	11.9(9.8)	12.0(8.0)	12.2(7.7)	12.4(7.3)	7.0(0.4)
3200	21.2(10.2)	20.7(9.4)	21.8(11.0)	20.7(9.4)	20.4(8.8)	20.3(8.5)	7.0(0.3)

4.1.2 Multiple index model

We next consider another model misspecification setting that involves the multiple index model

$$Y = f(\beta_1 X_1) + f(\beta_2 X_2 + \beta_3 X_3) + f(\beta_4 X_4 + \beta_5 X_5) + \varepsilon, \quad (29)$$

where the response depends on the covariates only through the first five ones but with non-linear functions and $f(x) = x^3/(x^2 + 1)$. Here the design matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p)$ was generated as in Section 4.1.1. We set the true parameter vector $\boldsymbol{\beta}_0 = (1, 1, -1, 1, -1, 0, \dots, 0)^T$, $n = 100$, and $\sigma = 0.25$. Note that the oracle working model is $M_0 = \text{supp}(\boldsymbol{\beta}_0) = \{1, \dots, 5\}$ for this example. Although the data was generated from model (29), we fit the linear regression model (2). The results are summarized in Tables 3 and 8 (the latter available in Supplementary Material). The consistent selection probability and inclusion probability are now calculated based on M_0 .

In general, the conclusions are similar to those in Example 4.1.1. An interesting observation is the comparison between $\text{GBIC}_p\text{-L}$ and GBIC_p in terms of model selection. While $\text{GBIC}_p\text{-L}$ is comparable to GBIC_p when the dimension is not large ($p = 200$), the difference between these two methods increases as the dimensionality increases. In the case when $p = 3200$, GBIC_p has 77% success probability of consistent selection, while all the other criteria have at most 5% success probability. This confirms the necessity of including the $\log p$ factor in the model selection criterion to take into account the high dimensionality, which is in line with the conclusion in Fan and Tang (2013) for the case of correctly specified models.

4.1.3 Logistic regression with interaction

Our last simulation example is high-dimensional logistic regression with interaction. We simulated 100 data sets from the logistic regression model with interaction and an n -dimensional parameter vector

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + 2\mathbf{x}_{p+1} + 2\mathbf{x}_{p+2}, \quad (30)$$

where $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p)$ is an $n \times p$ design matrix, $\mathbf{x}_{p+1} = \mathbf{x}_1 \circ \mathbf{x}_2$ and $\mathbf{x}_{p+2} = \mathbf{x}_3 \circ \mathbf{x}_4$ are two interaction terms, and the rest is the same as in (28). For each data set, the n -dimensional response vector \mathbf{y} was sampled from the Bernoulli distribution with success

Table 3: Simulation results for Example 4.1.2 with all entries multiplied by 100.

Consistent selection probability with inclusion probability							
p	AIC	BIC	GAIC	GBIC	GBIC _{p} -L	GBIC _{p}	Oracle
200	2(100)	4(100)	2(100)	6(100)	51(100)	65(100)	100(100)
400	1(100)	1(100)	2(100)	1(100)	28(100)	67(100)	100(100)
1600	0(100)	0(100)	3(100)	0(100)	5(100)	63(100)	100(100)
3200	0(100)	0(100)	5(100)	0(100)	5(100)	77(100)	100(100)
Median prediction error with RSD in parentheses							
200	26(3)	26(3)	26(3)	26(3)	23(3)	23(2)	22(1)
400	28(3)	28(3)	27(3)	28(3)	25(4)	23(2)	22(1)
1600	31(3)	31(3)	30(4)	31(3)	30(4)	23(4)	22(1)
3200	31(4)	31(4)	30(3)	31(4)	30(3)	23(2)	22(1)

probability vector $[e^{\theta_1}/(1 + e^{\theta_1}), \dots, e^{\theta_n}/(1 + e^{\theta_n})]^T$ with $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)^T$ given in (30). As in Section 4.1.1, we consider the case where all covariates are independent of each other. We chose $\boldsymbol{\beta}_0 = (2.5, -1.9, 2.8, -2.2, 3, 0, \dots, 0)^T$ and set sample size $n = 200$. Although the data was generated from the logistic regression model with parameter vector (30), we fit the logistic regression model without the two interaction terms. This provides another example of misspecified models. As argued in Section 4.1.1, the oracle working model is $\text{supp}(\boldsymbol{\beta}_0) = \{1, \dots, 5\}$ which corresponds to the logistic regression model with the first five covariates.

Since the goal in logistic regression is usually classification, we replace the prediction error with the classification error rate. Tables 4 and 9 (the latter available in Supplementary Material) show similar phenomenon as in Sections 4.1.1 and 4.1.2. Again GBIC _{p} outperformed all other model selection criteria with greater advantage for the high-dimensional case (e.g., $p = 3200$).

Table 4: Simulation results for Example 4.1.3 with all entries multiplied by 100.

Consistent selection probability with inclusion probability							
p	AIC	BIC	GAIC	GBIC	GBIC _{p} -L	GBIC _{p}	Oracle
200	0(99)	32(94)	1(99)	39(94)	49(91)	49(91)	100(100)
400	0(99)	19(97)	0(99)	36(93)	50(92)	55(92)	100(100)
1600	0(96)	0(96)	0(94)	21(90)	35(88)	47(81)	100(100)
3200	0(95)	0(95)	0(96)	10(90)	21(86)	41(72)	100(100)
Median classification error rate with RSD in parentheses							
200	22(3)	15(2)	16(2)	15(1)	14(1)	14(1)	14(1)
400	21(3)	16(5)	17(2)	15(1)	15(1)	15(1)	13(1)
1600	21(2)	21(2)	18(1)	16(3)	15(1)	16(2)	14(1)
3200	22(2)	21(2)	19(2)	18(3)	15(2)	15(2)	13(1)

4.2 Real data examples

We finally consider two gene expression data sets: Prostate (Singh et al., 2002) and Neuroblastoma (Oberthuer et al., 2006). The prostate data set contains $p = 12601$ genes with $n = 136$ samples including 59 positives and 77 negatives. The neuroblastoma (NB) data set, available from the MicroArray Quality Control phase-II (MAQC-II) project (MAQC Consortium, 2010), consists of gene expression profiles for $p = 10707$ genes from 239 patients (49 positives and 190 negatives) of the German Neuroblastoma Trials NB90-NB2004 with the 3-year event-free survival (3-year EFS) information available. See those references for more detailed description of the data sets.

We fit the logistic regression model with SICA implemented with ICA algorithm (Fan and Lv, 2011). Before applying the regularization method, we exploited the sure independence screening approach to reduce the dimensionality. The random permutation idea (Fan et al., 2011) was applied to determine the threshold for marginal screening. After the screening step, the numbers of retained variables are 430 (prostate) and 2778 (neuroblastoma), respectively. We then chose the final model using those six model selection criteria. Moreover, we randomly split the data into training (80%) and testing (20%) sets for 100 times, and reported the

Table 5: Results for Prostate and Neuroblastoma data sets.

Median classification error rate (in percentage) with RSD in parentheses						
Data set	AIC	BIC	GAIC	GBIC	GBIC _p -L	GBIC _p
Prostate	19(9)	15(6)	15(9)	15(9)	13(9)	15(10)
NB	18(5)	18(5)	18(3)	18(3)	18(5)	19(5)
Median model size with RSD in parentheses						
Prostate	15.0(3.7)	8.5(4.5)	3.0(1.5)	6.0(3.7)	6.0(3.7)	5.0(3.0)
NB	27.0(3.0)	26.0(2.2)	8.5(3.7)	6.0(3.7)	5.0(3.0)	3.0(2.2)

median test classification error rate along with the median model size in Table 5.

From Table 5, for the prostate data set the best criterion appears to be GBIC_p-L, which has the smallest test classification error rate. For the neuroblastoma data set, if we only look at the median test classification error rate, GBIC_p-L again has the best performance with a small model size. It is worth noting that GBIC_p leads to the most parsimonious model, with median model size 3, at the expense of slightly increasing the test classification error rate. From the results of real examples, it is evident that by taking into account the effect of model misspecification, the performance of the original model selection criteria can be improved in general. This is important since the true model structure is generally unavailable to us in real applications. Our results suggest that the term involving model misspecification in the asymptotic expansions is usually nonnegligible for model selection.

5 Discussion

Despite the rich literature on model selection, the general case of model misspecification in high dimensions is less well studied. Our work has investigated the problem of model selection in high-dimensional misspecified models and characterized the impact of model misspecification. The newly suggested information criterion GBIC_p involving a logarithmic factor of the dimensionality in penalizing model complexity has been shown to perform well in high-dimensional settings. Moreover, we have established the consistency of the covariance contrast matrix estimator that captures the effect of model misspecification in the general

setting.

The $\log p$ term in GBIC_p is adaptive to high dimensions. In the setting of correctly specified models, Fan and Tang (2013) showed that such a term is necessary for the model selection consistency of information criteria when the dimensionality diverges fast with the sample size. It would be interesting to study the optimality of those different information criteria under model misspecification. It would also be interesting to investigate model selection principles in more general high-dimensional misspecified models such as the additive models and survival models. These problems are beyond the scope of the current paper and are interesting topics for future research.

A Proofs of some main results

This appendix presents the proofs of Theorems 1 and 3–4. To save space, the proofs of all other theorems and technical lemmas are included in the Supplementary Material. For notational simplicity, throughout the proofs we may specify the orders of different quantities without stating the exact constants, and use the notation \mathbf{y} for observed response and \mathbf{Y} for random response interchangeably when it is convenient.

A.1 Proof of Theorem 1

Throughout this proof $\|\cdot\|_2$ denotes the Euclidean norm of a given vector. The main idea of the proof is to obtain a probabilistic lower bound for the event $\{\widehat{\boldsymbol{\beta}}_n \in N_n(\delta_n)\}$. To accomplish that we first consider an event that is a subset of this event and calculate the probabilistic lower bound for the smaller event.

Recall that $b''(\theta)$ is continuous and bounded away from 0, and \mathbf{X} is of full column rank d . Recall the definition $N_n(\delta_n) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|(n^{-1}\mathbf{B}_n)^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_{n,0})\|_2 \leq (n/d)^{-1/2}\delta_n\}$, and let $\partial N_n(\delta_n)$ denote the boundary of this neighborhood. Since $N_n(\delta_n)$ is compact, $\ell_n(\mathbf{y}, \cdot)$ a continuous strictly concave function, whenever the event

$$Q_n = \left\{ \ell_n(\mathbf{y}, \boldsymbol{\beta}_{n,0}) > \max_{\boldsymbol{\beta} \in \partial N_n(\delta_n)} \ell_n(\mathbf{y}, \boldsymbol{\beta}) \right\} \quad (31)$$

occurs, $\widehat{\boldsymbol{\beta}}_n$ will be in $N_n(\delta_n)$. The strict concavity of the log-likelihood function follows from the positive definiteness of $\mathbf{A}_n(\boldsymbol{\beta}) = \mathbf{X}^T \boldsymbol{\Sigma}(\mathbf{X}\boldsymbol{\beta})\mathbf{X}$, which is the negative of the Hessian of the

log-likelihood. This property entails that on the event Q_n , the global maximizer $\widehat{\beta}_n$ must belong to the interior of the neighborhood $N_n(\delta_n)$. Hereafter we condition on the event Q_n defined in (31). The technical arguments that follow herein, in order to prove that Q_n holds with significant probability, require delicate analyses due to growing dimensionality d .

Applying Taylor's expansion to the log-likelihood function $\ell_n(\mathbf{y}, \cdot)$ around $\beta_{n,0}$, we obtain

$$\ell_n(\mathbf{y}, \beta) - \ell_n(\mathbf{y}, \beta_{n,0}) = (\beta - \beta_{n,0})^T \Psi_n(\beta_{n,0}) - \frac{1}{2}(\beta - \beta_{n,0})^T \mathbf{A}_n(\beta_*)(\beta - \beta_{n,0}),$$

where β_* is on the line segment joining β and $\beta_{n,0}$ and $\Psi_n(\beta_{n,0}) = \mathbf{X}^T[\mathbf{y} - \mu(\mathbf{X}\beta_{n,0})]$. By letting $\mathbf{u} = \delta_n^{-1}d^{-1/2}\mathbf{B}_n^{1/2}(\beta - \beta_{n,0})$, the above Taylor's expansion can be rewritten as

$$\ell_n(\mathbf{y}, \beta) - \ell_n(\mathbf{y}, \beta_{n,0}) = d^{1/2}\delta_n \mathbf{u}^T \mathbf{B}_n^{-1/2} \Psi_n(\beta_{n,0}) - d\delta_n^2 \mathbf{u}^T \mathbf{V}_n(\beta_*) \mathbf{u} / 2, \quad (32)$$

where $\mathbf{V}_n(\beta) = \mathbf{B}_n^{-1/2} \mathbf{A}_n(\beta) \mathbf{B}_n^{-1/2}$.

From the definition of \mathbf{u} , $\beta \in \partial N_n(\delta_n)$ is equivalent to $\|\mathbf{u}\|_2 = 1$, and $\beta \in \partial N_n(\delta_n)$ implies $\beta_* \in N_n(\delta_n)$ since $N_n(\delta_n)$ is convex. Also it is clear that

$$\max_{\|\mathbf{u}\|_2=1} \mathbf{u}^T \mathbf{B}_n^{-1/2} \Psi_n(\beta_{n,0}) = \|\mathbf{B}_n^{-1/2} \Psi_n(\beta_{n,0})\|_2. \quad (33)$$

From Condition 2, for n sufficiently large, $\min_{\beta \in N_n(\delta_n)} \lambda_{\min} \{\mathbf{V}_n(\beta)\} > c_1 n^{-r}$ where $0 < r < 1/4$. Using this condition and since $\beta_* \in N_n(\delta_n)$, it holds that

$$\min_{\|\mathbf{u}\|_2=1} \mathbf{u}^T \mathbf{V}_n(\beta_*) \mathbf{u} \geq \min_{\beta \in N_n(\delta_n)} \lambda_{\min} \{\mathbf{V}_n(\beta)\} > c_1 n^{-r}. \quad (34)$$

Hence by combining (33)–(34) and taking a supremum on the boundary $\partial N_n(\delta_n)$ in (32) we derive

$$\begin{aligned} \max_{\beta \in \partial N_n(\delta_n)} \ell_n(\mathbf{y}, \beta) - \ell_n(\mathbf{y}, \beta_{n,0}) &< d^{1/2}\delta_n [\|\mathbf{B}_n^{-1/2} \Psi_n(\beta_{n,0})\|_2 \\ &\quad - 2^{-1}c_1 n^{-r} d^{1/2}\delta_n]. \end{aligned} \quad (35)$$

By (7), we have $\mathbf{X}^T[E\mathbf{y} - \mu(\mathbf{X}\beta_{n,0})] = 0$. Hence, $\Psi_n(\beta_{n,0}) = \mathbf{X}^T[\mathbf{y} - \mu(\mathbf{X}\beta_{n,0})] = \mathbf{X}^T(\mathbf{y} - E\mathbf{y})$. Denote by $\mathbf{W} = \mathbf{B}_n^{-1/2} \Psi_n(\beta_{n,0}) = \mathbf{B}_n^{-1/2} \mathbf{X}^T(\mathbf{y} - E\mathbf{y})$. Notice that $E\mathbf{W} = \mathbf{0}$ and $\text{cov}(\mathbf{W}) = \mathbf{B}_n^{-1/2} \text{cov}(\Psi_n(\beta_{n,0})) \mathbf{B}_n^{-1/2} = \mathbf{B}_n^{-1/2} \mathbf{B}_n \mathbf{B}_n^{-1/2} = I_d$.

Clearly the left hand side of (35) is negative with probability given by

$$P\{\|\mathbf{W}\|_2 \leq 2^{-1}c_1 n^{-r} d^{1/2}\delta_n\}. \quad (36)$$

From the expression of \mathbf{W} , we have

$$\begin{aligned}\|\mathbf{W}\|_2^2 &= (\mathbf{y} - E\mathbf{y})^T \mathbf{X} \mathbf{B}_n^{-1} \mathbf{X}^T (\mathbf{y} - E\mathbf{y}) \\ &= [(\mathbf{y} - E\mathbf{y})^T \text{cov}(\mathbf{y})^{-1/2}] [\text{cov}(\mathbf{y})^{1/2} \mathbf{X} \mathbf{B}_n^{-1} \mathbf{X}^T \text{cov}(\mathbf{y})^{1/2}] \\ &\quad \cdot [\text{cov}(\mathbf{y})^{-1/2} (\mathbf{y} - E\mathbf{y})],\end{aligned}$$

where \cdot denotes product. Denote by $\mathbf{R} = \text{cov}(\mathbf{y})^{1/2} \mathbf{X} \mathbf{B}_n^{-1} \mathbf{X}^T \text{cov}(\mathbf{y})^{1/2}$ and $\mathbf{q} = \text{cov}(\mathbf{y})^{-1/2} (\mathbf{y} - E\mathbf{y})$. It is easy to check that $\mathbf{R}^2 = \mathbf{R}$. Therefore, \mathbf{R} is a projection matrix with rank $\text{tr}(\mathbf{R}) = d$. In addition, we have $E\mathbf{q} = \mathbf{0}$ and $\text{cov}(\mathbf{q}) = I_n$.

We now decompose $\|\mathbf{W}\|_2^2$ into two terms, the summations of the diagonal entries and the off-diagonal entries, respectively,

$$\|\mathbf{W}\|_2^2 = \sum_{i=1}^n r_{ii} q_i^2 + \sum_{1 \leq i \neq j \leq n} r_{ij} q_i q_j, \quad (37)$$

where r_{ij} denotes the (i, j) -entry of \mathbf{R} . Next we obtain probabilistic bounds for each of the two terms.

From the sub-Gaussian tail condition for \mathbf{q} in Condition 1, there exists some positive constant H such that for any $t \geq 0$,

$$P(|q_i| > t) \leq H \exp(-t^2/H), \quad (38)$$

for all $1 \leq i \leq n$.

Thus for any $t \geq 0$, it holds that

$$P\left\{\bigcap_{i=1}^n \{q_i^2 \leq t^2\}\right\} \geq 1 - \sum_{i=1}^n P\{q_i^2 > t^2\} \geq 1 - nH \exp(-t^2/H). \quad (39)$$

On the event $\bigcap_{i=1}^n \{q_i^2 \leq t^2\}$, we can bound the first term of (37) as

$$\sum_{i=1}^n r_{ii} q_i^2 \leq t^2 \text{tr}(\mathbf{R}) = dt^2. \quad (40)$$

Denote by \mathbf{R}_D a diagonal matrix with diagonal entries r_{ii} . As a result, we observe that $\sum_{1 \leq i \neq j \leq n} r_{ij} q_i q_j = \mathbf{q}^T (\mathbf{R} - \mathbf{R}_D) \mathbf{q}$. It is easy to see that $E[\mathbf{q}^T (\mathbf{R} - \mathbf{R}_D) \mathbf{q}] = 0$. We will use a version of the Hanson-Wright inequality (see, e.g., Theorem 1.1 of Rudelson and Vershynin, 2013) to obtain the concentration bound of the quadratic form $\mathbf{q}^T (\mathbf{R} - \mathbf{R}_D) \mathbf{q}$. But we first start with some notation and preparation.

Let $\|\xi\|_{\psi_2}$ denote the sub-Gaussian norm of a sub-Gaussian random variable ξ defined as $\|\xi\|_{\psi_2} = \sup_{m \geq 1} \{m^{-1/2}(E|\xi|^m)^{1/m}\}$. From Condition 1, that is, the condition on sub-Gaussian tails, we derive

$$\begin{aligned} E|q_i|^m &= m \int_0^\infty x^{m-1} P(|q_i| \geq x) dx \leq Hm \int_0^\infty x^{m-1} \exp(-x^2/H) dx \\ &= (Hm/2) H^{m/2} \int_0^\infty u^{m/2-1} \exp(-u) du \\ &= (Hm/2) H^{m/2} \Gamma(m/2) \leq (Hm/2) H^{m/2} (m/2)^{m/2}, \end{aligned}$$

where the last line follows directly from the definition of the Gamma function. Taking the m -th root, we have

$$(E|q_i|^m)^{1/m} \leq (Hm/2)^{1/m} H^{1/2} (m/2)^{1/2}.$$

Rewriting after bounding $(1/2)^{(1/m)+(1/2)}$ by 1, we obtain

$$m^{-1/2} (E|q_i|^m)^{1/m} \leq m^{1/m} H^{(1/2)+(1/m)} \leq e^{1/e} (H^{3/2} \vee 1)$$

since $m \geq 1$. Therefore, it holds that $\|q_i\|_{\psi_2} \leq c_4$ for all i , where $c_4 = e^{1/e} (H^{3/2} \vee 1)$.

We now need bounds on the operator and Frobenius norms of $\mathbf{R} - \mathbf{R}_D$. Denote $\|\cdot\|_2$ and $\|\cdot\|_F$ as the matrix operator and Frobenius norms, respectively. Note that $\|\mathbf{R}\|_2 = 1$ and $\|\mathbf{R}\|_F^2 = \text{tr}(\mathbf{R}^2) = \text{tr}(\mathbf{R}) = d$. Thus using the fact $\sum_{i \neq j} r_{ij}^2 \leq d$, we obtain $\|\mathbf{R} - \mathbf{R}_D\|_F^2 \leq d$. Since $|r_{ii}| \leq 1$, we further obtain $\|\mathbf{R} - \mathbf{R}_D\|_2 \leq \|\mathbf{R}\|_2 + \|\mathbf{R}_D\|_2 \leq 2$. Thereby, a direct application of the Hanson-Wright inequality yields

$$\begin{aligned} P \left\{ \left| \sum_{i \neq j} r_{ij} q_i q_j \right| > dt^2 \right\} &= P \{ |\mathbf{q}^T (\mathbf{R} - \mathbf{R}_D) \mathbf{q}| > dt^2 \} \\ &\leq 2 \exp \left\{ -c_5 \min \left(\frac{d^2 t^4}{c_4^4 \|\mathbf{R} - \mathbf{R}_D\|_F^2}, \frac{dt^2}{c_4^2 \|\mathbf{R} - \mathbf{R}_D\|_2} \right) \right\} \\ &\leq 2 \exp \left\{ -c_5 \min \left(\frac{dt^4}{c_4^4}, \frac{dt^2}{2c_4^2} \right) \right\} \\ &\leq 2 \exp \{-c_6 dt^2\} \end{aligned} \tag{41}$$

for any $t > c_4/\sqrt{2}$, where c_5 and c_6 are some positive constants. To ensure $t > c_4/\sqrt{2}$, we choose $\delta_n = n^r (c_0 \log n)^{1/2}$ for some constant $c_0 > 8c_1^{-2}H$ and $t = 2^{-3/2}c_1 n^{-r} \delta_n$. Therefore, the probability bound (41) holds for large enough n .

Combining (39) and (41), with probability at least $1 - nH \exp(-t^2/H) - 2 \exp(-c_6 dt^2)$, we have

$$\|\mathbf{W}\|_2^2 \leq \sum_i r_{ii} q_i^2 + \left| \sum_{i \neq j} r_{ij} q_i q_j \right| 2dt^2.$$

In view of our choice of t , it holds that $(2dt^2)^{1/2} = 2^{-1} c_1 n^{-r} d^{1/2} \delta_n$ and thus

$$\begin{aligned} P\{\|\mathbf{W}\|_2 \leq 2^{-1} c_1 n^{-r} d^{1/2} \delta_n\} &\geq 1 - nH \exp(-t^2/H) - 2 \exp(-c_6 dt^2) \\ &\geq 1 - O(n^{-\alpha}), \end{aligned}$$

where $\alpha = [(c_1^2 c_0 / (8H) - 1)] \wedge (c_1^2 c_6 c_0 / 8)$. Note that $\alpha > 0$ since $c_0 > 8c_1^{-2} H$. This leads to

$$P(Q_n) \geq 1 - O(n^{-\alpha}). \quad (42)$$

The positive constant α can be large if c_0 in δ_n is chosen to be large. From Condition 2, $\lambda_{\min}(\mathbf{B}_n) \rightarrow \infty$ at a faster rate than $d\delta_n^2$. Then we have the consistency $\hat{\beta}_n - \beta_{n,0} = o_P(1)$.

A.2 Proof of Theorem 3

Define $\mathcal{E} = \{\hat{\beta}_n \in N_n(\delta_n)\}$, where $\hat{\beta}_n$ stands for the QMLE. Note that \mathcal{E} does depend on n , but for simplicity of notation we will omit the subscript n in sequel. To establish this theorem we require a possibly dimension dependent bound on the quantity $\|n^{-1/2} \mathbf{X} \hat{\beta}_n\|_2$. The need for bounding the specified quantity, particularly with growing dimensionality, can be intuitively understood by trying to put some restriction on the parameter space. This is analogous to the case of penalized likelihood.

Recall the neighborhood $M_n(\alpha_1) = \{\beta \in \mathbb{R}^d : \|\mathbf{X}\beta\|_\infty \leq \alpha_1 \log n\}$, where α_1 is some positive constant satisfying $\alpha_1 < \alpha/2 - 1$. One way of bounding the quantity $\|n^{-1/2} \mathbf{X} \hat{\beta}_n\|_2$ is to restrict the QMLE $\hat{\beta}_n$ on the set $M_n(\alpha_1)$. As mentioned in Theorem 1, the constant α can be large if c_0 is chosen to be large, which ensures that α_1 is positive. From Condition 5, $N_n(\delta_n) \subset M_n(\alpha_1)$ for all sufficiently large n to ensure that conditional on \mathcal{E} , the restricted MLE coincides with its unrestricted version. However, this condition is very mild in the sense that the constant α_1 can be chosen as large as desired to make $M_n(\alpha_1)$ large enough, whereas the neighborhood $N_n(\delta_n)$ is asymptotically shrinking. Hereafter in this proof $\hat{\beta}_n$ will be referred to as the restricted MLE, unless specified otherwise.

Recall that $\eta_n(\boldsymbol{\beta}) = E\ell_n(\tilde{\mathbf{y}}, \boldsymbol{\beta})$, where $\tilde{\mathbf{y}}$ is an independent copy of \mathbf{y} . In the GLM setup, we have $\ell_n(\tilde{\mathbf{y}}, \boldsymbol{\beta}) = \tilde{\mathbf{y}}^T \mathbf{X}\boldsymbol{\beta} - \mathbf{1}^T \mathbf{b}(\mathbf{X}\boldsymbol{\beta})$ and $\eta_n(\boldsymbol{\beta}) = (E\tilde{\mathbf{y}}^T) \mathbf{X}\boldsymbol{\beta} - \mathbf{1}^T \mathbf{b}(\mathbf{X}\boldsymbol{\beta})$.

Part 1: Expansion of $E\eta_n(\hat{\boldsymbol{\beta}}_n)$. We approach the proof by splitting $E\eta_n(\hat{\boldsymbol{\beta}}_n)$ in the region \mathcal{E} and its complement, that is,

$$\begin{aligned} E\eta_n(\hat{\boldsymbol{\beta}}_n) &= E\{\eta_n(\hat{\boldsymbol{\beta}}_n)1_{\mathcal{E}}\} + E\{\eta_n(\hat{\boldsymbol{\beta}}_n)1_{\mathcal{E}^c}\} \\ &= E\{\eta_n(\hat{\boldsymbol{\beta}}_n)1_{\mathcal{E}}\} + E\{[(E\tilde{\mathbf{y}})^T(\mathbf{X}\hat{\boldsymbol{\beta}}_n) - \mathbf{1}^T \mathbf{b}(\mathbf{X}\hat{\boldsymbol{\beta}}_n)]1_{\mathcal{E}^c}\}, \end{aligned} \quad (43)$$

where the second equality follows from the definition of $\eta_n(\cdot)$.

We aim to show that the second term on the right hand side of (43) is $o(1)$. Performing componentwise Taylor's expansion of $\mathbf{b}(\cdot)$ around $\mathbf{0}$ and evaluating at $\mathbf{X}\hat{\boldsymbol{\beta}}_n$, we obtain $\mathbf{b}(\mathbf{X}\hat{\boldsymbol{\beta}}_n) = \mathbf{b}(\mathbf{0}) + b'(\mathbf{0})\mathbf{X}\hat{\boldsymbol{\beta}}_n + \mathbf{r}$, where $\mathbf{r} = (r_1, \dots, r_n)^T$ with $r_i = 2^{-1}b''((\mathbf{X}\boldsymbol{\beta}_i^*)_i)(\mathbf{X}\hat{\boldsymbol{\beta}}_n)_i^2$ and $\boldsymbol{\beta}_1^*, \dots, \boldsymbol{\beta}_n^*$ lying in the line segment joining $\hat{\boldsymbol{\beta}}_n$ and $\mathbf{0}$. Recall that $\hat{\boldsymbol{\beta}}_n$ is the constrained MLE here, EY_i^2 is bounded uniformly in i and n , and $b''(\cdot) = O(n^{\alpha_1})$ uniformly in its argument. The condition on $b''(\cdot)$ can be much weakened in many cases including linear and logistic regression models. This condition also accommodates Poisson regression where $b''(\theta) = \exp(\theta)$ for $\theta \in \mathbb{R}$ since $b(\theta) = \exp(\theta)$. Then it follows that

$$\begin{aligned} E\{|(E\tilde{\mathbf{y}})^T \mathbf{X}\hat{\boldsymbol{\beta}}_n - \mathbf{1}^T \mathbf{b}(\mathbf{X}\hat{\boldsymbol{\beta}}_n)|1_{\mathcal{E}^c}\} &\leq O\{n \log n + n + n^{1+\alpha_1}(\log n)^2\}P(\mathcal{E}^c) \\ &\leq O\{n^{2(\alpha_1+1)}\}P(\mathcal{E}^c) = o(1) \end{aligned} \quad (44)$$

for sufficiently large n . The last inequality follows from the fact that $\alpha > 2(\alpha_1 + 1)$ and we recall that $P(\mathcal{E}^c) = O(n^{-\alpha})$. To verify the orders, we note that the four bounds $|(E\tilde{\mathbf{y}})^T \mathbf{X}\hat{\boldsymbol{\beta}}_n| \leq n \max_{1 \leq i \leq n} (EY_i^2)^{1/2} \alpha_1 \log n$, $|\mathbf{1}^T \mathbf{b}(\mathbf{0})| = O(n)$, $|b'(\mathbf{0})\mathbf{1}^T \mathbf{X}\hat{\boldsymbol{\beta}}_n| \leq O(1)n\alpha_1 \log n$, and $|\mathbf{1}^T \mathbf{r}| \leq n \max_{1 \leq i \leq n} |r_i| \leq nO(n^{\alpha_1})(\alpha_1 \log n)^2$.

On the event \mathcal{E} , we first expand $\eta_n(\boldsymbol{\beta})$ around $\boldsymbol{\beta}_{n,0}$. By the definition of $\boldsymbol{\beta}_{n,0}$, $\eta_n(\boldsymbol{\beta})$ attains its maximum at $\boldsymbol{\beta}_{n,0}$. By Taylor's expansion of $\eta_n(\cdot)$ around $\boldsymbol{\beta}_{n,0}$ and evaluating at $\hat{\boldsymbol{\beta}}_n$, we derive

$$\begin{aligned} \eta_n(\hat{\boldsymbol{\beta}}_n) &= \eta_n(\boldsymbol{\beta}_{n,0}) - \frac{1}{2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0})^T \mathbf{A}_n(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0}) \\ &= \eta_n(\boldsymbol{\beta}_{n,0}) - \frac{1}{2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0})^T \mathbf{A}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0}) - \frac{s_n}{2} \\ &= \eta_n(\boldsymbol{\beta}_{n,0}) - \frac{1}{2}\mathbf{v}_n^T[(\mathbf{C}_n^{-1})^T \mathbf{A}_n \mathbf{C}_n^{-1}]\mathbf{v}_n - \frac{s_n}{2}, \end{aligned} \quad (45)$$

where $\mathbf{A}_n(\cdot) = -\partial^2 \ell_n(\mathbf{y}, \cdot) / \partial \beta^2$, $\mathbf{A}_n = \mathbf{A}_n(\beta_{n,0})$, $s_n = (\hat{\beta}_n - \beta_{n,0})^T [\mathbf{A}_n(\beta^*) - \mathbf{A}_n](\hat{\beta}_n - \beta_{n,0})$, $\mathbf{v}_n = \mathbf{C}_n(\hat{\beta}_n - \beta_{n,0})$, and β^* is on the line segment joining $\beta_{n,0}$ and $\hat{\beta}_n$. Then it follows that

$$\begin{aligned} |s_n 1_{\mathcal{E}}| &= \left| (\hat{\beta}_n - \beta_{n,0})^T (\mathbf{A}_n(\beta^*) - \mathbf{A}_n)(\hat{\beta}_n - \beta_{n,0}) \right| 1_{\mathcal{E}} \\ &= \left| [\mathbf{B}_n^{1/2}(\hat{\beta}_n - \beta_{n,0})]^T [\mathbf{V}_n(\beta^*) - \mathbf{V}_n] [\mathbf{B}_n^{1/2}(\hat{\beta}_n - \beta_{n,0})] \right| 1_{\mathcal{E}} \\ &\leq \|\mathbf{V}_n(\beta^*) - \mathbf{V}_n\|_2 \delta_n^2 d 1_{\mathcal{E}}, \end{aligned} \quad (46)$$

where $\mathbf{V}_n(\cdot) = \mathbf{B}_n^{-1/2} \mathbf{A}_n(\cdot) \mathbf{B}_n^{-1/2}$ and $\mathbf{V}_n = \mathbf{V}_n(\beta_{n,0})$. Note that on the event \mathcal{E} , by the convexity of the neighborhood $N_n(\delta_n)$ we have $\beta^* \in N_n(\delta_n)$. From Condition 4, $\max_{\beta_1, \dots, \beta_d \in N_n(\delta_n)} \|\tilde{\mathbf{V}}_n(\beta_1, \dots, \beta_d) - \mathbf{V}_n\|_2 = O(d^{1/2} n^{-1/2})$. Therefore we deduce that $E(s_n 1_{\mathcal{E}})$ is of order $O(d^{3/2} n^{-1/2} \delta_n^2) = o(1)$, which follows from (A.3) in the proof of Theorem 2.

From (A.2) in the proof of Theorem 2, we have the decomposition $\mathbf{v}_n = \mathbf{u}_n + \mathbf{w}_n$ with $\mathbf{u}_n = \mathbf{B}_n^{-1/2} \mathbf{X}^T(\mathbf{y} - E\mathbf{y})$ and

$$\mathbf{w}_n = - \left[\tilde{\mathbf{V}}_n(\beta_1, \dots, \beta_d) - \mathbf{V}_n \right] \left[\mathbf{B}_n^{1/2}(\hat{\beta}_n - \beta_{n,0}) \right].$$

For simplicity of notation, denote by $\mathbf{R}_n = (\mathbf{C}_n^{-1})^T \mathbf{A}_n \mathbf{C}_n^{-1}$. Recall that $\mathbf{C}_n = \mathbf{B}_n^{-1/2} \mathbf{A}_n$. With some calculations we obtain

$$\begin{aligned} E(\mathbf{u}_n^T \mathbf{R}_n \mathbf{u}_n) &= E\{(\mathbf{y} - E\mathbf{y})^T \mathbf{X} \mathbf{A}_n^{-1} \mathbf{X}^T (\mathbf{y} - E\mathbf{y})\} \\ &= E\{\text{tr}(\mathbf{A}_n^{-1} \mathbf{X}^T (\mathbf{y} - E\mathbf{y})(\mathbf{y} - E\mathbf{y})^T \mathbf{X})\} = \text{tr}(\mathbf{A}_n^{-1} \mathbf{B}_n). \end{aligned}$$

Note that $E(\mathbf{u}_n^T \mathbf{R}_n \mathbf{u}_n 1_{\mathcal{E}}) = E(\mathbf{u}_n^T \mathbf{R}_n \mathbf{u}_n) - E(\mathbf{u}_n^T \mathbf{R}_n \mathbf{u}_n 1_{\mathcal{E}^c})$. From Theorem 1, we have $P(\mathcal{E}^c) \rightarrow 0$ as $n \rightarrow \infty$. Let $\mu_n = \text{tr}(\mathbf{A}_n^{-1} \mathbf{B}_n) \vee 1$ ensuring that this quantity is bounded away from zero. We will apply Vitali's convergence theorem to show that $E(\mathbf{u}_n^T \mathbf{R}_n \mathbf{u}_n 1_{\mathcal{E}^c}) = o(\mu_n)$. To establish uniform integrability we use the following lemma, the proof of which has been provided in Appendix C in Supplementary Material.

Lemma 1. *For some constant $\gamma > 0$, $\sup_n E|(\mathbf{u}_n^T \mathbf{R}_n \mathbf{u}_n) / \mu_n|^{1+\gamma} < \infty$.*

This leads to $E(\mathbf{u}_n^T \mathbf{R}_n \mathbf{u}_n 1_{\mathcal{E}^c}) = o(\mu_n)$. Hence we have

$$\frac{1}{2} E(\mathbf{u}_n^T \mathbf{R}_n \mathbf{u}_n 1_{\mathcal{E}}) = \frac{1}{2} \text{tr}(\mathbf{A}_n^{-1} \mathbf{B}_n) + o(\mu_n).$$

It remains to show that

$$E[(\mathbf{w}_n^T \mathbf{R}_n \mathbf{w}_n + 2\mathbf{w}_n^T \mathbf{R}_n \mathbf{u}_n) 1_{\mathcal{E}}] = o(\mu_n). \quad (47)$$

Note that on the event \mathcal{E} , we have

$$\mathbf{w}_n^T \mathbf{R}_n \mathbf{w}_n = \|\mathbf{R}_n^{1/2} \mathbf{w}_n\|_2^2 \leq \|\tilde{\mathbf{V}}_n - \mathbf{V}_n\|_2^2 \delta_n^2 d \text{tr}(\mathbf{A}_n^{-1} \mathbf{B}_n).$$

In view of the assumption $\max_{\beta_1, \dots, \beta_d \in N_n(\delta_n)} \|\tilde{\mathbf{V}}_n(\beta_1, \dots, \beta_d) - \mathbf{V}_n\|_2 = O(d^{1/2} n^{-1/2})$, it holds that $E(\mathbf{w}_n^T \mathbf{R}_n \mathbf{w}_n | \mathcal{E}) = o(\mu_n)$. For the cross term $\mathbf{w}_n^T \mathbf{R}_n \mathbf{u}_n$, applying the Cauchy-Schwarz inequality yields

$$\begin{aligned} |E(\mathbf{w}_n^T \mathbf{R}_n \mathbf{u}_n | \mathcal{E})| &\leq E(\|\mathbf{R}_n^{1/2} \mathbf{w}_n\|_2^2 | \mathcal{E})^{1/2} E(\|\mathbf{u}_n^T \mathbf{R}_n^{1/2}\|_2^2 | \mathcal{E})^{1/2} \\ &\leq E[\|\tilde{\mathbf{V}}_n - \mathbf{V}_n\|_2 | \mathcal{E} \delta_n d^{1/2} \text{tr}(\mathbf{A}_n^{-1} \mathbf{B}_n)], \end{aligned}$$

which entails that $E(\mathbf{w}_n^T \mathbf{R}_n \mathbf{u}_n | \mathcal{E}) = o(\mu_n)$. Note that $E\{|\eta_n(\beta_{n,0})| | \mathcal{E}^c\}$ is of order $o(1)$ by similar calculations as in (44). Thus combining (43) – (47) yields $E\{\eta_n(\hat{\beta}_n)\} = \eta_n(\beta_{n,0}) - \frac{1}{2} \text{tr}(\mathbf{A}_n^{-1} \mathbf{B}_n) + o(\mu_n)$.

Part 2: Expansion of $E\ell_n(\mathbf{y}, \beta_{n,0})$. Similarly we expand $\ell_n(\mathbf{y}, \cdot)$ around $\hat{\beta}_n$ and evaluate at $\beta_{n,0}$. From Condition 5, $N_n(\delta_n) \subset M_n(\alpha_1)$ for sufficiently large n , we see that $\beta_{n,0} \in M_n(\alpha_1)$. On the event \mathcal{E} , since $\ell_n(\mathbf{y}, \cdot)$ attains its maximum at the restricted MLE $\hat{\beta}_n$, we have

$$\begin{aligned} \ell_n(\mathbf{y}, \beta_{n,0}) &= \ell_n(\mathbf{y}, \hat{\beta}_n) - \frac{1}{2} (\hat{\beta}_n - \beta_{n,0})^T \mathbf{A}_n(\beta^*) (\hat{\beta}_n - \beta_{n,0}) \\ &= \ell_n(\mathbf{y}, \hat{\beta}_n) - \frac{1}{2} (\hat{\beta}_n - \beta_{n,0})^T \mathbf{A}_n(\hat{\beta}_n) (\hat{\beta}_n - \beta_{n,0}) - \frac{s_n}{2} \\ &= \ell_n(\mathbf{y}, \hat{\beta}_n) - \frac{1}{2} \mathbf{v}_n^T [(\mathbf{C}_n^{-1})^T \mathbf{A}_n \mathbf{C}_n^{-1}] \mathbf{v}_n - \frac{s_n}{2}. \end{aligned} \tag{48}$$

Then similarly as in Part 1, we can obtain

$$E\{\ell_n(\mathbf{y}, \beta_{n,0}) | \mathcal{E}\} = E\{\ell_n(\mathbf{y}, \hat{\beta}_n) | \mathcal{E}\} - \frac{1}{2} \text{tr}(\mathbf{A}_n^{-1} \mathbf{B}_n) + o(\mu_n).$$

If we can show that $E\{|\ell_n(\mathbf{y}, \beta_{n,0})| | \mathcal{E}^c\}$ and $E\{|\ell_n(\mathbf{y}, \hat{\beta}_n)| | \mathcal{E}^c\}$ are both of order $o(1)$, then we obtain the desired asymptotic expansion

$$E\{\eta_n(\hat{\beta}_n)\} = E\{\ell_n(\mathbf{y}, \hat{\beta}_n)\} - \text{tr}(\mathbf{A}_n^{-1} \mathbf{B}_n) + o(\mu_n).$$

To see why $E\{|\ell_n(\mathbf{y}, \beta_{n,0})| | \mathcal{E}^c\}$ is of order $o(1)$, we derive

$$\begin{aligned} &E\{|\mathbf{y}^T \mathbf{X} \beta_{n,0} - \mathbf{1}^T \mathbf{b}(\mathbf{X} \beta_{n,0})| | \mathcal{E}^c\} \\ &\leq O(n \log n) P(\mathcal{E}^c)^{1/2} + O\{n + n \log n + n^{2+\alpha_1} (\log n)^2\} P(\mathcal{E}^c) \\ &\leq O\{n^{2(\alpha_1+1)}\} P(\mathcal{E}^c) = o(1), \end{aligned}$$

similarly as in (44) and using $E[|\mathbf{y}^T \mathbf{X} \boldsymbol{\beta}_{n,0}|1_{\mathcal{E}^c}] \leq E[|\mathbf{y}^T \mathbf{X} \boldsymbol{\beta}_{n,0}|^2]^{1/2} P(\mathcal{E}^c)^{1/2}$. Similarly we can also show that $E\{|\ell_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n)|1_{\mathcal{E}^c}\}$ is of order $o(1)$. The only difference in the above derivation is to bound $\|\mathbf{X} \hat{\boldsymbol{\beta}}_n\|_\infty$ instead of $\|\mathbf{X} \boldsymbol{\beta}_{n,0}\|_\infty$, which holds from the definition of the restricted QMLE. This concludes the proof.

A.3 Proof of Theorem 4

In view of the expansions of GAIC, GBIC, and GBIC_p, we need to show that $\log |\hat{\mathbf{H}}_n| = \log |\mathbf{H}_n| + o_P(1)$ and $\text{tr}(\hat{\mathbf{H}}_n) = \text{tr}(\mathbf{H}_n) + o_P(1)$. To establish this we show that $\hat{\mathbf{H}}_n = \mathbf{H}_n + o_P(1/d)$, where the $o_P(\cdot)$ denotes the convergence in probability of the matrix operator norm.

Let \mathbf{M} be a $d \times d$ square matrix. Denote by $\overline{\text{tr}}(\mathbf{M}) = \text{tr}(\mathbf{M})/d$ the normalized trace and $\rho(\mathbf{M}) = \max_{1 \leq k \leq d} \{|\lambda_k(\mathbf{M})|\}$ the spectral radius. Then we have

$$\begin{aligned} |\text{tr}(\hat{\mathbf{H}}_n) - \text{tr}(\mathbf{H}_n)| &= d |\overline{\text{tr}}(\hat{\mathbf{H}}_n - \mathbf{H}_n)| \\ &\leq d \rho(\hat{\mathbf{H}}_n - \mathbf{H}_n) = d \|\hat{\mathbf{H}}_n - \mathbf{H}_n\|_2 = o_P(1), \end{aligned}$$

where $\|\cdot\|_2$ denotes the matrix operator norm. The equality of the spectral radius and the operator norm follows from the symmetry of the matrix $\hat{\mathbf{H}}_n - \mathbf{H}_n$. Similarly define the normalized log determinant, that is, $\overline{\log} |\mathbf{M}| = (\log |\mathbf{M}|)/d$ for any arbitrary matrix \mathbf{M} . Denote $\lambda_k(\cdot)$ as the eigenvalues arranged in the increasing order. Then we have

$$\begin{aligned} |\log |\hat{\mathbf{H}}_n| - \log |\mathbf{H}_n|| &\leq d |\overline{\log} |\hat{\mathbf{H}}_n| - \overline{\log} |\mathbf{H}_n|| \\ &\leq d \max_{1 \leq k \leq d} |\log \lambda_k(\hat{\mathbf{H}}_n) - \log \lambda_k(\mathbf{H}_n)| \\ &\leq d \max_{1 \leq k \leq d} \log \left\{ 1 + \left| \frac{\lambda_k(\hat{\mathbf{H}}_n)}{\lambda_k(\mathbf{H}_n)} - 1 \right| \right\}. \end{aligned} \tag{49}$$

Recall that we assume that the smallest and largest eigenvalues of both $n^{-1} \mathbf{B}_n$ and $n^{-1} \mathbf{A}_n$ are bounded away from 0 and ∞ . It then follows that $\lambda_k(\mathbf{H}_n) = O(1)$ and $\lambda_k^{-1}(\mathbf{H}_n) = O(1)$ uniformly for all $1 \leq k \leq d$. An application of Weyl's theorem shows that

$$|\lambda_k(\hat{\mathbf{H}}_n) - \lambda_k(\mathbf{H}_n)| \leq \rho(\hat{\mathbf{H}}_n - \mathbf{H}_n)$$

for each k . We have $\rho(\hat{\mathbf{H}}_n - \mathbf{H}_n) = \|\hat{\mathbf{H}}_n - \mathbf{H}_n\|_2 = o_P(1/d)$. Hence the right hand side of (49) is $o_P(1)$.

Now we proceed to show that $\widehat{\mathbf{H}}_n = \mathbf{H}_n + o_P(1/d)$. It suffices to prove that $n^{-1}\widehat{\mathbf{A}}_n = n^{-1}\mathbf{A}_n + o_P(1/d)$ and $n^{-1}\widehat{\mathbf{B}}_n = n^{-1}\mathbf{B}_n + o_P(1/d)$. We use the following properties of the operator norm (Horn and Johnson, 1985): $\|(I_d - \mathbf{M})^{-1}\|_2 \leq 1/(1 - \|\mathbf{M}\|_2)$ if $\|\mathbf{M}\|_2 < 1$, $\|\mathbf{M}\mathbf{N}\|_2 \leq \|\mathbf{M}\|_2\|\mathbf{N}\|_2$, and $\|\mathbf{M} + \mathbf{N}\|_2 \leq \|\mathbf{M}\|_2 + \|\mathbf{N}\|_2$, where \mathbf{M} and \mathbf{N} are $d \times d$ matrices. To see the sufficiency note that

$$\begin{aligned} & (n^{-1}\widehat{\mathbf{A}}_n)^{-1}(n^{-1}d\widehat{\mathbf{B}}_n) - (n^{-1}\mathbf{A}_n)^{-1}(n^{-1}d\mathbf{B}_n) \\ &= (n^{-1}\widehat{\mathbf{A}}_n)^{-1}(n^{-1}d\widehat{\mathbf{B}}_n) - (n^{-1}\widehat{\mathbf{A}}_n)^{-1}(n^{-1}d\mathbf{B}_n) + (n^{-1}\widehat{\mathbf{A}}_n)^{-1}(n^{-1}d\mathbf{B}_n) \\ & \quad - (n^{-1}\mathbf{A}_n)^{-1}(n^{-1}d\mathbf{B}_n). \end{aligned}$$

Then the desired result $\widehat{\mathbf{H}}_n = \mathbf{H}_n + o_P(1/d)$ can be obtained by repeated application of the above properties of the operator norm.

Part 1: Prove $n^{-1}\widehat{\mathbf{A}}_n = n^{-1}\mathbf{A}_n + o_P(1/d)$. From Theorem 1 we have, $\|(n^{-1}\mathbf{B}_n)^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0})\|_2 = O_P\{(n/d)^{-1/2}\delta_n\}$, which along with the assumption that the smallest eigenvalue of $n^{-1}\mathbf{B}_n$ is bounded away from 0 entails $\widehat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_{n,0} + O_P\{(n/d)^{-1/2}\delta_n\}$. Then it follows from the Lipschitz assumption for $n^{-1}\mathbf{A}_n(\boldsymbol{\beta})$ from Condition 7 in the neighborhood $N_n(\delta_n)$ and Theorem 1 that $n^{-1}\widehat{\mathbf{A}}_n = n^{-1}\mathbf{A}_n + o_P(1/d)$, which holds for our choice of $d = o\{n^{(1-4r)/3}(\log n)^{-2/3}\}$ and δ_n .

Part 2: Prove $n^{-1}\widehat{\mathbf{B}}_n = n^{-1}\mathbf{B}_n + o_P(1/d)$. We first split $n^{-1}\widehat{\mathbf{B}}_n$ as

$$n^{-1}\widehat{\mathbf{B}}_n = n^{-1}\mathbf{X}^T \text{diag} \left\{ \left[\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}}_n) \right] \circ \left[\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}}_n) \right] \right\} \mathbf{X} = \mathbf{G}_1 + \mathbf{G}_2 + \mathbf{G}_3,$$

where

$$\begin{aligned} \mathbf{G}_1 &= n^{-1}\mathbf{X}^T \text{diag} \{ (\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})) \circ (\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})) \} \mathbf{X}, \\ \mathbf{G}_2 &= 2n^{-1}\mathbf{X}^T \text{diag} \{ (\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})) \circ [\boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0}) - \boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}}_n)] \} \mathbf{X}, \\ \mathbf{G}_3 &= n^{-1}\mathbf{X}^T \text{diag} \{ [\boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}}_n) - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})] \circ [\boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}}_n) - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})] \} \mathbf{X}. \end{aligned}$$

We will state two lemmas before proceeding with the proof. Define the sub-exponential norm of a sub-exponential random variable ξ as

$$\|\xi\|_{\psi_1} = \sup_{m \geq 1} \left\{ m^{-1} (E|\xi|^m)^{1/m} \right\}.$$

Lemma 2. For independent sub-Gaussian random variables $\{y_i\}_{i=1}^n$, we have that $q_i^2 = (y_i - Ey_i)^2 / \text{var}(y_i)$ is sub-exponential with norm bounded by $2c_4^2$, where c_4 is as defined in the proof of Theorem 1. Moreover, the following Bernstein-type tail probability bound holds

$$P\{|\sum_{i=1}^n a_i q_i^2 - E[\sum_{i=1}^n a_i q_i^2]| \geq t\} \leq 2 \exp \left[-c_{10} \min \left(\frac{t^2}{4c_4^4 \|\mathbf{a}\|_2^2}, \frac{t}{2c_4^2 \|\mathbf{a}\|_\infty} \right) \right]$$

for $\mathbf{a} \in \mathbb{R}^n$, $t \geq 0$, and $c_{10} > 0$.

Lemma 3. For independent sub-Gaussian random variables $\{y_i\}_{i=1}^n$ with $q_i = \{\text{var}(y_i)\}^{-1/2} (y_i - Ey_i)$, the following tail probability bound holds

$$P\{|\sum_{i=1}^n a_i q_i| \geq t\} \leq e \exp \left(-\frac{c_{11} t^2}{c_4^2 \|\mathbf{a}\|_2^2} \right)$$

for $\mathbf{a} \in \mathbb{R}^n$, $t \geq 0$, and $c_{11} > 0$.

Lemma 2 follows from Lemma 5.14 and Proposition 5.16 of Vershynin (2012). Note that here we define the sub-exponential random variable as the square of a sub-Gaussian random variable and the bound on the norm follows by our previous observation that $\|q_i\|_{\Psi_2} \leq c_4$ in the proof of Theorem 1. Lemma 3 rephrases Proposition 5.10 of Vershynin (2012) for the case where $\|q_i\|_{\Psi_2} \leq c_4$.

Further split \mathbf{G}_1 as $\mathbf{G}_1 = \mathbf{G}_{11} + \mathbf{G}_{12} + \mathbf{G}_{13}$ where

$$\begin{aligned} \mathbf{G}_{11} &= n^{-1} \mathbf{X}^T \text{diag}\{(\mathbf{y} - E\mathbf{y}) \circ (\mathbf{y} - E\mathbf{y})\} \mathbf{X}, \\ \mathbf{G}_{12} &= 2n^{-1} \mathbf{X}^T \text{diag}\{(\mathbf{y} - E\mathbf{y}) \circ [E\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})]\} \mathbf{X}, \\ \mathbf{G}_{13} &= n^{-1} \mathbf{X}^T \text{diag}\{[E\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})] \circ [E\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})]\} \mathbf{X}. \end{aligned}$$

Note that $E\mathbf{G}_{11} = n^{-1} \mathbf{B}_n$ and $\mathbf{G}_{11} = n^{-1} \sum_{i=1}^n \{\mathbf{x}_i \mathbf{x}_i^T [y_i - Ey_i]^2\} = \sum_{i=1}^n \mathbf{A}_i q_i^2$, where $\mathbf{A}_i = n^{-1} \text{var}(y_i) \mathbf{x}_i \mathbf{x}_i^T$. Then it holds that for any positive t ,

$$\begin{aligned} P(\|\mathbf{G}_{11} - E\mathbf{G}_{11}\|_2 \geq t) &\leq P(\|\mathbf{G}_{11} - E\mathbf{G}_{11}\|_F \geq t) \\ &\leq d^2 \max_{1 \leq j, k \leq d} P(|\mathbf{G}_{11}^{jk} - E\mathbf{G}_{11}^{jk}| \geq t/d), \end{aligned} \quad (50)$$

where $\|\cdot\|_F$ denotes the matrix Frobenius norm and \mathbf{G}_{11}^{jk} denotes the (j, k) entry of \mathbf{G}_{11} . Recall from Condition 7 that $\|\mathbf{X}\|_\infty = O(n^{\alpha_2})$ with $0 \leq \alpha_2 < r$. Define $a_i^{jk} = n^{-1} \text{var}(y_i) x_{ij} x_{ik}$ and $\mathbf{a}^{jk} = (a_1^{jk}, \dots, a_n^{jk})^T$. We have $\|\mathbf{a}^{jk}\|_2^2 = O(n^{-1} n^{4\alpha_2})$. Then combining (50) with

Lemma 2, we deduce

$$\begin{aligned} P(d\|\mathbf{G}_{11} - E\mathbf{G}_{11}\|_2 \geq t) &\leq d^2 \max_{1 \leq j, k \leq d} P(|\mathbf{G}_{11}^{jk} - E\mathbf{G}_{11}^{jk}| \geq t/d^2) \\ &\leq 2d^2 \exp\{-c_{12}t^2 n^{1-4\alpha_2}/d^4\} \end{aligned}$$

for some constant $c_{12} > 0$. Note that $d = o\{n^{(1-4r)/4}\}$, we obtain $\mathbf{G}_{11} = E\mathbf{G}_{11} + o_P(1/d)$.

By Condition 8 and Lemma 3, we have

$$P(d\|\mathbf{G}_{12}\|_2 \geq t) \leq d^2 P(|\mathbf{G}_{12}^{jk}| \geq t/d^2) \leq ed^2 \exp\{-c_{13}t^2 n^{1-4\alpha_2+(1-\alpha_3)/2}/d^4\},$$

where $c_{13} > 0$ is some constant. Hence from $d = o\{n^{(1-4r)/4}\}$ and $0 \leq \alpha_3 \leq 4(r - \alpha_2)$, we have $\mathbf{G}_{12} = o_P(1/d)$.

To show that $\mathbf{G}_{13} = o(1/d)$, we derive

$$\begin{aligned} \|\mathbf{G}_{13}\|_2^2 &\leq \|n^{-1}\sum_{i=1}^n \{\mathbf{x}_i \mathbf{x}_i^T [Ey_i - [\boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})]_i]^2\}\|_F^2 \\ &= \sum_{1 \leq j, k \leq d} [\sum_{i=1}^n a_i^{jk} [Ey_i - [\boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})]_i]^2 / \text{var}(y_i)]^2 \\ &\leq \sum_{i=1}^n \{[Ey_i - [\boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})]_i]^2 / \text{var}(y_i)\}^2 \sum_{1 \leq j, k \leq d} \|\mathbf{a}^{jk}\|_2^2, \end{aligned}$$

where the last step follows from the component-wise Cauchy-Schwarz inequality. From Condition 8, $\mathbf{G}_{13} = o(1/d)$. Combining the above derivations yields $\mathbf{G}_1 = E\mathbf{G}_1 + o_P(1/d) = n^{-1}\mathbf{B}_n + o_P(1/d)$. To see that $\mathbf{G}_2 = o_P(1/d)$, note that $(\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0}))_i = (y_i - Ey_i) + (Ey_i - [\boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})]_i)$ and apply similar arguments as above. By the Lipschitz Condition 7 in the neighborhood $N_n(\delta_n)$, we have $\mathbf{G}_3 = o_P(1/d)$, which completes the proof.

References

- [1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (eds. B. N. Petrov and F. Csaki), Akademiai Kiado, Budapest, 267–281.
- [2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Auto. Control* **19**, 716–723.
- [3] Bozdogan, H. (1987). Model selection and Akaike’s information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52**, 345–370.

- [4] Chang, C.-H., Huang, H.-C. and Ing, C.-K. (2014). Asymptotic theory of generalized information criterion for geostatistical regression model selection, *Ann. Statist.*, to appear.
- [5] Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759–771.
- [6] Chen, K. and Chan, K. S. (2011). Subset ARMA model selection via the adaptive Lasso. *Statistics and Its Interface* **4**, 197–205.
- [7] Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *J. Amer. Statist. Assoc.* **106**, 544–557.
- [8] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.
- [9] Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory* **57**, 5467–5484.
- [10] Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *J. Roy. Statist. Soc. Ser. B* **75**, 531–552.
- [11] Foster, D. and George, E. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947–1975.
- [12] Hall, P. (1990). Akaike’s information criterion and Kullback-Leibler loss for histogram density estimation. *Probability Theory and Related Fields* **85**, 449–467.
- [13] Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis*. New York: Cambridge University Press.
- [14] Ing, C.-K. (2007). Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *Ann. Statist.* **35**, 1238–1277.
- [15] Ing, C.-K. and Lai, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica* **21**, 1473–1513.

- [16] Konishi, S. and Kitagawa, G. (1996). Generalised information criterion in model selection. *Biometrika* **83**, 875–890.
- [17] Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86.
- [18] Liu, W. and Yang, Y. (2011). Parametric or nonparametric? A parametricness index for model selection. *Ann. Statist.* **39**, 2074–2102.
- [19] Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37**, 3498–3528.
- [20] Lv, J. and Liu, J. S. (2014). Model selection principles in misspecified models. *J. Roy. Statist. Soc. Ser. B* **76**, 141–167.
- [21] MAQC Consortium (2010). The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.* **28**, 827–841.
- [22] Oberthuer, A., Berthold, F., Warnat, P., Hero, B., Kahlert, Y., Spitz, R., Ernestus, K., König, R., Haas, S., Eils, R., Schwab, M., Brors, B., Westermann, F. and Fischer, M. (2006) Customized oligonucleotide microarray gene expression based classification of neuroblastoma patients outperforms current clinical risk stratification. *Journal of Clinical Oncology* **24**, 5070–5078.
- [23] Peng, H., Yan, H. and Zhang, W. (2013). The connection between cross-validation and Akaike information criterion in a semiparametric family. *Journal of Nonparametric Statistics* **25**, 475–485.
- [24] Rudelson, M. and Vershynin, R. (2013). Hanson-Wright inequality and sub-gaussian concentration. *Electron. Commun. Probab.* **18**, 1–9.
- [25] Schwartz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.
- [26] Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., Damico, A. and Richie, J. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203–209.

- [27] Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc. Ser. B* **39**, 44–47.
- [28] Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing, Theory and Applications* **Chapter 5**, 210–268. Cambridge University Press.
- [29] Wang, H., Li, R. and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–568.
- [30] Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Statist. Soc. B* **71**, 671–683.
- [31] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.
- [32] Zhang, Y., Li, R. and Tsai, C. L. (2010). Regularization parameter selections via generalized information criterion *J. Amer. Statist. Assoc.* **105**, 312–323.

Supplementary Material to “Model Selection in High-Dimensional Misspecified Models”

Pallavi Basu, Yang Feng and Jinchi Lv

This Supplementary Material contains the proofs of Theorems 2 and 5–6, and technical lemmas, as well as additional tables from Section 4.1.

B Proofs of Additional Theorems

B.1 Proof of Theorem 2

Recall that $\mathbf{C}_n = \mathbf{B}_n^{-1/2} \mathbf{A}_n$. To establish the asymptotic normality of the QMLE $\hat{\beta}_n$, we prove the following

$$\mathbf{D}_n \mathbf{C}_n (\hat{\beta}_n - \beta_{n,0}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, I_m), \quad (\text{A.1})$$

for any $m \times d$ matrix \mathbf{D}_n such that $\mathbf{D}_n \mathbf{D}_n^T = I_m$ with m fixed. From the score equation we have $\Psi(\hat{\beta}_n) = \mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\hat{\beta}_n)] = 0$. From (7), it holds that $\mathbf{X}^T[E\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\beta_{n,0})] = 0$. For any $\beta_1, \dots, \beta_d \in \mathbb{R}^d$, denote by $\tilde{\mathbf{A}}_n(\beta_1, \dots, \beta_d)$ a $d \times d$ matrix with j -th row the corresponding row of $\mathbf{A}_n(\beta_j)$ for each $j = 1, \dots, d$, and matrix-valued function $\tilde{\mathbf{V}}_n(\beta_1, \dots, \beta_d) = \mathbf{B}_n^{-1/2} \tilde{\mathbf{A}}_n(\beta_1, \dots, \beta_d) \mathbf{B}_n^{-1/2}$. Assuming the differentiability of $\Psi(\cdot)$ and applying the mean-value theorem componentwise around $\beta_{n,0}$, we obtain

$$\begin{aligned} \mathbf{0} &= \Psi_n(\hat{\beta}_n) = \Psi_n(\beta_{n,0}) - \tilde{\mathbf{A}}_n(\beta_1, \dots, \beta_d)(\hat{\beta}_n - \beta_{n,0}) \\ &= \mathbf{X}^T(\mathbf{y} - E\mathbf{y}) - \tilde{\mathbf{A}}_n(\beta_1, \dots, \beta_d)(\hat{\beta}_n - \beta_{n,0}), \end{aligned}$$

where each of β_1, \dots, β_d lies on the line segment joining $\hat{\beta}_n$ and $\beta_{n,0}$. It follows from this expansion that

$$\mathbf{C}_n(\hat{\beta}_n - \beta_{n,0}) = \mathbf{u}_n + \mathbf{w}_n, \quad (\text{A.2})$$

where $\mathbf{u}_n = \mathbf{B}_n^{-1/2} \mathbf{X}^T(\mathbf{y} - E\mathbf{y})$ and

$$\mathbf{w}_n = - \left[\tilde{\mathbf{V}}_n(\beta_1, \dots, \beta_d) - \mathbf{V}_n \right] \left[\mathbf{B}_n^{1/2}(\hat{\beta}_n - \beta_{n,0}) \right],$$

where $\mathbf{V}_n = \mathbf{V}_n(\beta_{n,0}) = \mathbf{B}_n^{-1/2} \mathbf{A}_n \mathbf{B}_n^{-1/2}$. Therefore we have

$$\mathbf{D}_n \mathbf{C}_n(\hat{\beta}_n - \beta_{n,0}) = \mathbf{D}_n \mathbf{u}_n + \mathbf{D}_n \mathbf{w}_n.$$

By the Cramér-Wold theorem, it suffices to show that for any unit vector $\mathbf{a} \in \mathbb{R}^m$, $\mathbf{a}^T \mathbf{D}_n \mathbf{C}_n(\hat{\beta}_n - \beta_{n,0}) \xrightarrow{\mathcal{D}} N(0, 1)$. Further by Slutsky's lemma, it is sufficient to show that $\mathbf{a}^T \mathbf{D}_n \mathbf{u}_n \xrightarrow{\mathcal{D}} N(0, 1)$ and $\mathbf{a}^T \mathbf{D}_n \mathbf{w}_n = o_P(1)$ for any unit vector \mathbf{a} .

Part 1 (Asymptotic normality of $\mathbf{a}^T \mathbf{D}_n \mathbf{u}_n$): We will build on the conditions required to apply the Lyapunov central limit theorem (CLT). For an arbitrary unit vector $\mathbf{a} \in \mathbb{R}^m$, consider the asymptotic distribution of

$$v_n = \mathbf{a}^T \mathbf{D}_n \mathbf{u}_n = \mathbf{a}^T \mathbf{D}_n \mathbf{B}_n^{-1/2} \mathbf{X}^T (\mathbf{y} - E\mathbf{y}) = \sum_{i=1}^n z_i,$$

where $z_i = \mathbf{a}^T \mathbf{D}_n \mathbf{B}_n^{-1/2} \mathbf{x}_i (y_i - Ey_i)$, $i = 1, \dots, n$, and $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$. Since z_i 's are independent and have mean zero, we derive

$$\begin{aligned} \text{var}(v_n) &= \sum_{i=1}^n \text{var}(z_i) = \mathbf{a}^T \mathbf{D}_n \mathbf{B}_n^{-1/2} \mathbf{X}^T \text{cov}(\mathbf{y}) \mathbf{X} \mathbf{B}_n^{-1/2} \mathbf{D}_n^T \mathbf{a} \\ &= \mathbf{a}^T \mathbf{D}_n \mathbf{B}_n^{-1/2} \mathbf{B}_n \mathbf{B}_n^{-1/2} \mathbf{D}_n^T \mathbf{a} = 1. \end{aligned}$$

From Condition 3, we have $\max_{1 \leq i \leq n} E|y_i - Ey_i|^3 \leq M$ for some positive constant M and $\sum_{i=1}^n (\mathbf{x}_i^T \mathbf{B}_n^{-1} \mathbf{x}_i)^{3/2} = o(1)$. Then an application of the Cauchy-Schwarz inequality yields

$$\begin{aligned} \sum_{i=1}^n E|z_i|^3 &= \sum_{i=1}^n |\mathbf{a}^T \mathbf{D}_n \mathbf{B}_n^{-1/2} \mathbf{x}_i|^3 E|y_i - Ey_i|^3 \leq M \sum_{i=1}^n |\mathbf{a}^T \mathbf{D}_n \mathbf{B}_n^{-1/2} \mathbf{x}_i|^3 \\ &\leq M \sum_{i=1}^n \|\mathbf{D}_n^T \mathbf{a}\|_2^3 \|\mathbf{B}_n^{-1/2} \mathbf{x}_i\|_2^3 = M \sum_{i=1}^n (\mathbf{x}_i^T \mathbf{B}_n^{-1} \mathbf{x}_i)^{3/2} \rightarrow 0, \end{aligned}$$

noting that $\|\mathbf{D}_n^T \mathbf{a}\|_2^2 = \mathbf{a}^T \mathbf{D}_n \mathbf{D}_n^T \mathbf{a} = \mathbf{a}^T I_m \mathbf{a} = 1$. Therefore by applying Lyapunov's CLT, we obtain

$$\mathbf{a}^T \mathbf{D}_n \mathbf{u}_n = \sum_{i=1}^n z_i \xrightarrow{\mathcal{D}} N(0, 1).$$

Part 2 (To show $\mathbf{a}^T \mathbf{D}_n \mathbf{w}_n$ is $o(1)$ in probability): Conditional on the event $\{\hat{\beta}_n \in N_n(\delta_n)\}$ and using the fact that $\|\mathbf{D}_n^T \mathbf{a}\|_2 = 1$, we have

$$\begin{aligned} |\mathbf{a}^T \mathbf{D}_n \mathbf{w}_n| &\leq \|\mathbf{D}_n^T \mathbf{a}\|_2 \|\mathbf{w}_n\|_2 \leq \|\mathbf{w}_n\|_2 \\ &\leq \|\tilde{\mathbf{V}}_n - \mathbf{V}_n\|_2 \|\mathbf{B}_n^{1/2}(\hat{\beta}_n - \beta_{n,0})\|_2 \\ &\leq \|\tilde{\mathbf{V}}_n - \mathbf{V}_n\|_2 d^{1/2} \delta_n, \end{aligned}$$

where the last step follows from the definition of the neighborhood $N_n(\delta_n) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|(n^{-1}\mathbf{B}_n)^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_{n,0})\|_2 \leq (n/d)^{-1/2}\delta_n\}$ and given that $\{\widehat{\boldsymbol{\beta}}_n \in N_n(\delta_n)\}$. From Condition 4, $\max_{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_d \in N_n(\delta_n)} \|\widetilde{\mathbf{V}}_n(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_d) - \mathbf{V}_n\|_2 = O(d^{1/2}n^{-1/2}) \leq O(dn^{-1/2}\delta_n)$. Again conditional on the event $\{\widehat{\boldsymbol{\beta}}_n \in N_n(\delta_n)\}$ and noticing that each $\boldsymbol{\beta}_j$ defined previously for $1 \leq j \leq d$ lies in $N_n(\delta_n)$ due to its convexity, it holds that

$$|\mathbf{a}^T \mathbf{D}_n \mathbf{w}_n| = O(d^{3/2}n^{-1/2}\delta_n^2) = o(1), \quad (\text{A.3})$$

where we choose $\delta_n = n^r(c_0 \log n)^{1/2}$ as in the proof of Theorem 1 and $d = o\{n^{(1-4r)/3}(\log n)^{-2/3}\}$ with $0 \leq r < 1/4$. Since the event $\{\widehat{\boldsymbol{\beta}}_n \in N_n(\delta_n)\}$ holds with probability tending to 1, $\mathbf{a}^T \mathbf{D}_n \mathbf{w}_n = o_P(1)$. Also note that the convergence to zero in probability is uniform in \mathbf{a} and \mathbf{D}_n . Therefore, combining parts 1 and 2 finishes the proof.

B.2 Proof of Theorem 5

Throughout the proof we condition on the event $\widetilde{Q}_n = \{\widehat{\boldsymbol{\beta}}_n \in N_n(\delta_n)\}$, where $N_n(\delta_n) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|(n^{-1}\mathbf{B}_n)^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_{n,0})\|_2 \leq (n/d)^{-1/2}\delta_n\}$, $\mathbf{B}_n = \mathbf{X}^T \text{cov}(\mathbf{Y}) \mathbf{X}$, and $\widehat{\boldsymbol{\beta}}_n$ is the unrestricted MLE. From Theorem 1 we have shown that as $n \rightarrow \infty$,

$$P(\widetilde{Q}_n) \rightarrow 1.$$

Recall from (20) that $\ell_n^*(\mathbf{y}, \boldsymbol{\beta}) = \ell_n(\mathbf{y}, \boldsymbol{\beta}) - \ell_n(\mathbf{y}, \widehat{\boldsymbol{\beta}}_n)$. Then the maximum value zero of this function is attained at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_n$. It follows from (9) that

$$\partial^2 \ell_n^*(\mathbf{y}, \boldsymbol{\beta}) / \partial \boldsymbol{\beta}^2 = -\mathbf{A}_n(\boldsymbol{\beta}),$$

where $\mathbf{A}_n(\boldsymbol{\beta}) = \mathbf{X}^T \boldsymbol{\Sigma}(\mathbf{X}\boldsymbol{\beta}) \mathbf{X}$. By Taylor's expansion of the likelihood function $\ell_n(\mathbf{y}, \cdot)$ around $\widehat{\boldsymbol{\beta}}_n$ in the new neighborhood $\widetilde{N}_n(\delta_n) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|(n^{-1}\mathbf{B}_n)^{1/2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n)\|_2 \leq (n/d)^{-1/2}\delta_n\}$, we derive

$$\begin{aligned} \ell_n^*(\mathbf{y}, \boldsymbol{\beta}) &= \frac{1}{2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n)^T [\partial^2 \ell_n^*(\mathbf{y}, \boldsymbol{\beta}_*) / \partial \boldsymbol{\beta}^2] (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n) \\ &= -\frac{n}{2} \boldsymbol{\delta}^T \mathbf{V}_n(\boldsymbol{\beta}_*) \boldsymbol{\delta}, \end{aligned} \quad (\text{A.4})$$

where $\boldsymbol{\beta}_*$ lies on the line segment joining $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}_n$, $\boldsymbol{\delta} = n^{-1/2}\mathbf{B}_n^{1/2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n)$, and $\mathbf{V}_n(\boldsymbol{\beta}) = \mathbf{B}_n^{-1/2} \mathbf{A}_n(\boldsymbol{\beta}) \mathbf{B}_n^{-1/2}$. Since $\widehat{\boldsymbol{\beta}}_n \in \widetilde{N}_n(\delta_n)$, by the convexity of the neighborhood $\widetilde{N}_n(\delta_n)$ we have

$\beta_* \in \tilde{N}_n(\delta_n)$. Also note that conditional on the event \tilde{Q}_n , it holds that $\tilde{N}_n(\delta_n) \subset N_n(2\delta_n)$.

We define

$$\rho_n(\delta_n) = \max_{\beta \in N_n(2\delta_n)} \max\{|\lambda_{\min}(\mathbf{V}_n(\beta) - \mathbf{V}_n)|, |\lambda_{\max}(\mathbf{V}_n(\beta) - \mathbf{V}_n)|\}$$

with $\mathbf{V}_n = \mathbf{V}_n(\beta_{n,0})$. Using Taylor's expansion (A.4) over the region $\tilde{N}_n(\delta_n)$, we obtain

$$q_1(\beta)1_{\tilde{N}_n(\delta_n)}(\beta) \leq -n^{-1}\ell_n^*(\mathbf{y}, \beta)1_{\tilde{N}_n(\delta_n)}(\beta) \leq q_2(\beta)1_{\tilde{N}_n(\delta_n)}(\beta), \quad (\text{A.5})$$

where $q_1(\beta) = \frac{1}{2}\delta^T[\mathbf{V}_n - \rho_n(\delta_n)I_d]\delta$ and $q_2(\beta) = \frac{1}{2}\delta^T[\mathbf{V}_n + \rho_n(\delta_n)I_d]\delta$.

Define $U_n(\beta) = \exp[n^{-1}\ell_n^*(\mathbf{y}, \beta)]$ which takes values in the interval $[0, 1]$ by definition. From Condition 2, for n large, $\min_{\beta \in N_n(\delta_n)} \lambda_{\min}\{\mathbf{V}_n(\beta)\} > c_1 n^{-r}$ with $0 < r < 1/4$ and $\rho_n(\delta_n) = o\{n^{-(1-r)/3}\}$. Since $\beta_{n,0}$ belongs to $N_n(\delta_n)$, this assumption yields $\rho_n(\delta_n) \leq \lambda_{\min}(\mathbf{V}_n)/2$ for sufficiently large n . To see this, note that since $(1-r)/3 > r$ we have $\rho_n(\delta_n)n^r = o(1)$ whereas $\lambda_{\min}(\mathbf{V}_n)n^r > c_1$. Consider the linear transformation $h(\beta) = (n^{-1}\mathbf{B}_n)^{1/2}\beta$. For sufficiently large n , we obtain

$$\begin{aligned} E_{\mu_{\mathfrak{M}}}[e^{-nq_2(\beta)}1_{\tilde{N}_n(\delta_n)}(\beta)] &\leq E_{\mu_{\mathfrak{M}}}[U_n(\beta)^n 1_{\tilde{N}_n(\delta_n)}(\beta)] \\ &\leq E_{\mu_{\mathfrak{M}}}[e^{-nq_1(\beta)}1_{\tilde{N}_n(\delta_n)}(\beta)], \end{aligned} \quad (\text{A.6})$$

where $\mu_{\mathfrak{M}}$ denotes the prior distribution on $h(\beta) \in \mathbb{R}^d$ for model \mathfrak{M} . Before proceeding with the proof we state a few useful lemmas. The proofs of these lemmas are elaborated in Appendix A. From Condition 6, the prior density relative to the Lebesgue measure μ_0 on \mathbb{R}^d , $\pi(h(\beta)) = \frac{d\mu_{\mathfrak{M}}}{d\mu_0}(h(\beta))$, satisfies

$$\inf_{\beta \in N_n(2\delta_n)} \pi(h(\beta)) \geq c_2 \text{ and } \sup_{\beta \in \mathbb{R}^d} \pi(h(\beta)) \leq c_3, \quad (\text{A.7})$$

where c_2 and c_3 are some positive constants.

Lemma 4. Under (A.7), for $j = 1, 2$, we have

$$c_2 \int_{\beta \in \mathbb{R}^d} e^{-nq_j} 1_{\tilde{N}_n(\delta_n)} d\mu_0 \leq E_{\mu_{\mathfrak{M}}}[e^{-nq_j} 1_{\tilde{N}_n(\delta_n)}] \leq c_3 \int_{\beta \in \mathbb{R}^d} e^{-nq_j} 1_{\tilde{N}_n(\delta_n)} d\mu_0. \quad (\text{A.8})$$

Lemma 5. Conditional on the event \tilde{Q}_n , for sufficiently large n we have

$$\begin{aligned} E_{\mu_{\mathfrak{M}}}[U_n(\beta)^n 1_{\tilde{N}_n(\delta_n)}] &\leq \exp\{-[\kappa_n - \rho_n(\delta_n)/2]d\delta_n^2\} \\ &\leq \exp[-(\kappa_n/2)d\delta_n^2], \end{aligned} \quad (\text{A.9})$$

where $\kappa_n = \lambda_{\min}(\mathbf{V}_n)/2$.

Lemma 6. *It holds that*

$$\int_{\boldsymbol{\delta} \in \mathbb{R}^d} e^{-nq_1} d\mu_0 = \left(\frac{2\pi}{n} \right)^{d/2} |\mathbf{V}_n - \rho_n(\delta_n) I_d|^{-1/2} \quad (\text{A.10})$$

and

$$\int_{\boldsymbol{\delta} \in \mathbb{R}^d} e^{-nq_2} d\mu_0 = \left(\frac{2\pi}{n} \right)^{d/2} |\mathbf{V}_n + \rho_n(\delta_n) I_d|^{-1/2}. \quad (\text{A.11})$$

Lemma 7. *For $j = 1, 2$, it holds that*

$$\int_{\boldsymbol{\delta} \in \mathbb{R}^d} e^{-nq_j} 1_{\tilde{N}_n^c(\delta_n)} d\mu_0 \leq \left(\frac{2\pi}{n\kappa_n} \right)^{d/2} \exp \left[-(\sqrt{\kappa_n d \delta_n^2} - \sqrt{d})^2 / 2 \right] \quad (\text{A.12})$$

Now we proceed with the proof. From Condition 2, $\delta_n = n^r (c_0 \log n)^{1/2}$. Then the expression in (A.9) converges to zero faster than any polynomial rate in n . Let us rewrite the right hand side of (A.12) as

$$\exp \left\{ -\frac{d}{2} (\sqrt{\kappa_n \delta_n^2} - 1)^2 + \frac{d}{2} [\log(2\pi) - \log(n\kappa_n)] \right\},$$

which converges to zero faster than any polynomial rate in n . From Condition 2, $d = o\{n^{(1-4r)/3} (\log n)^{-2/3}\}$ with $0 < r < 1/4$.

Then it follows that

$$\begin{aligned} |\mathbf{V}_n \pm \rho_n(\delta_n) I_d|^{-1/2} &= |\mathbf{V}_n|^{-1/2} |\mathbf{I}_d \pm \rho_n(\delta_n) \mathbf{V}_n^{-1}|^{-1/2} \\ &= |\mathbf{V}_n|^{-1/2} \{1 + O[\rho_n(\delta_n) \text{tr}(\mathbf{V}_n^{-1})]\} \\ &= |\mathbf{V}_n|^{-1/2} \{1 + O[\rho_n(\delta_n) d \lambda_{\min}^{-1}(\mathbf{V}_n)]\} \\ &= |\mathbf{V}_n|^{-1/2} [1 + o(1)]. \end{aligned}$$

Combining Lemmas 4–7 yields

$$\begin{aligned} \log E_{\mu_{\mathfrak{M}}}[U_n(\boldsymbol{\beta})^n] &= \log \left\{ \left(\frac{2\pi}{n} \right)^{d/2} |\mathbf{V}_n|^{-1/2} [1 + o(1)] \right\} + \log c_n \\ &= -\frac{\log n}{2} d + \frac{1}{2} \log |\mathbf{A}_n^{-1} \mathbf{B}_n| + \frac{\log(2\pi)}{2} d + \log c_n + o(1), \end{aligned}$$

where $c_n \in [c_2, c_3]$. This completes the proof.

B.3 Proof of Theorem 6

The proof follows from the proof of Theorem 5 and Part 1, that is, expansion of $E\eta_n(\hat{\boldsymbol{\beta}}_n)$, of the proof of Theorem 3.

C Proofs of Lemmas

Lemmas 2 and 3 have been discussed in the paragraph following them. The proofs of Lemmas 4-6 can be found in Lv and Liu (2014).

C.1 Proof of Lemma 1

From the expression of $\mathbf{u}_n^T \mathbf{R}_n \mathbf{u}_n$, we have

$$\begin{aligned} \mathbf{u}_n^T \mathbf{R}_n \mathbf{u}_n &= (\mathbf{y} - E\mathbf{y})^T \mathbf{X} \mathbf{A}_n^{-1} \mathbf{X}^T (\mathbf{y} - E\mathbf{y}) \\ &= [(\mathbf{y} - E\mathbf{y})^T \text{cov}(\mathbf{y})^{-1/2}] [\text{cov}(\mathbf{y})^{1/2} \mathbf{X} \mathbf{A}_n^{-1} \mathbf{X}^T \text{cov}(\mathbf{y})^{1/2}] \\ &\quad \cdot [\text{cov}(\mathbf{y})^{-1/2} (\mathbf{y} - E\mathbf{y})]. \end{aligned}$$

Denote by $\mathbf{S}_n = \text{cov}(\mathbf{y})^{1/2} \mathbf{X} \mathbf{A}_n^{-1} \mathbf{X}^T \text{cov}(\mathbf{y})^{1/2}$ and $\mathbf{q} = \text{cov}(\mathbf{y})^{-1/2} (\mathbf{y} - E\mathbf{y})$. We decompose $\mathbf{u}_n^T \mathbf{R}_n \mathbf{u}_n$ into two terms, the summations of the diagonal entries and the off-diagonal entries, respectively,

$$\mathbf{u}_n^T \mathbf{R}_n \mathbf{u}_n = \sum_{i=1}^n s_{ii} q_i^2 + \sum_{1 \leq i \neq j \leq n} s_{ij} q_i q_j,$$

where s_{ij} denotes the (i, j) -entry of \mathbf{S}_n . Then we have

$$\begin{aligned} E(\mathbf{u}_n^T \mathbf{R}_n \mathbf{u}_n)^2 &= \sum_{i=1}^n s_{ii}^2 E(q_i^4) + \sum_{1 \leq i \neq j \leq n} s_{ii} s_{jj} E(q_i^2) E(q_j^2) \\ &\quad + \sum_{1 \leq i \neq j \leq n} s_{ij}^2 E(q_i^2) E(q_j^2). \end{aligned}$$

Using the sub-Gaussian norm bound c_4 , both quantities $E(q_i^4)$ and $E(q_i^2)E(q_j^2)$ can be uniformly bounded by a common constant. Hence

$$E(\mathbf{u}_n^T \mathbf{R}_n \mathbf{u}_n)^2 \leq O(1) \cdot \{\text{tr}(\mathbf{S}_n)^2 + \text{tr}(\mathbf{S}_n^2)\}.$$

Since \mathbf{S}_n is positive semidefinite it holds that $\text{tr}(\mathbf{S}_n^2) \leq [\text{tr}(\mathbf{S}_n)]^2$. Finally noting that $\text{tr}(\mathbf{S}_n) = \text{tr}(\mathbf{A}_n^{-1} \mathbf{B}_n)$, we see that $\sup_n E|(\mathbf{u}_n^T \mathbf{R}_n \mathbf{u}_n) / \mu_n|^{1+\gamma} < \infty$ for $\gamma = 1$.

C.2 Proof of Lemma 7

From the definition of $q_j(\boldsymbol{\beta})$ for $j = 1, 2$, we derive

$$\begin{aligned}\exp(-nq_j) &= \exp(-(n/2)\boldsymbol{\delta}^T[\mathbf{V}_n \pm \rho_n(\delta_n)I_d]\boldsymbol{\delta}) \\ &\leq \exp(-n(\kappa_n - \rho_n(\delta_n)/2)\boldsymbol{\delta}^T\boldsymbol{\delta}) \\ &\leq \exp(-(n\kappa_n)/2\boldsymbol{\delta}^T\boldsymbol{\delta}).\end{aligned}\tag{A.13}$$

Then we have

$$\begin{aligned}\int_{\boldsymbol{\delta} \in \mathbb{R}^d} e^{-nq_j} 1_{\tilde{N}_n^c(\delta_n)} d\mu_0 &\leq \int_{\boldsymbol{\delta} \in \mathbb{R}^d} e^{-\frac{n\kappa_n}{2}\boldsymbol{\delta}^T\boldsymbol{\delta}} 1_{\tilde{N}_n^c(\delta_n)} d\mu_0 \\ &= \left(\frac{2\pi}{n\kappa_n}\right)^{d/2} P(\|(n\kappa_n)^{-1/2}\mathbf{Z}\|_2^2 \geq (n/d)^{-1}\delta_n^2) \\ &= \left(\frac{2\pi}{n\kappa_n}\right)^{d/2} P(\|\mathbf{Z}\|_2^2 \geq \kappa_n d \delta_n^2),\end{aligned}$$

where $\mathbf{Z} \sim N(\mathbf{0}, I_d)$.

Using the chi-square tail bound, that is, for any positive x it is known that $P(\|\mathbf{Z}\|_2^2 - d \geq 2\sqrt{dx} + 2x) \leq \exp(-x)$ and after minor modification it holds that $P(\|\mathbf{Z}\|_2^2 \geq (\sqrt{d} + \sqrt{2x})^2) \leq \exp(-x)$. With this observation, define $x = (\sqrt{\kappa_n d \delta_n^2} - \sqrt{d})^2/2$ and the proof concludes.

D Additional Tables

In Tables 6–9, we report additional variable selection results for the three simulation examples in Section 4.1.

Table 6: Example 4.1.1. Median false positives with median false negatives (strong/weak effects) in parentheses when the model is misspecified.

	AIC	BIC	GAIC	GBIC	GBIC _p -L	GBIC _p
200	24(0/4)	2(0/5)	2(0/5)	1(0/5)	0(0/5)	0(0/5)
400	23(0/4)	19(0/5)	3(0/5)	4(0/5)	0(0/5)	0(0/5)
1600	24(0/5)	25(0/5)	4(0/5)	23(0/5)	1(0/5)	0(0/5)
3200	19(0/5)	25(0/5)	4(0/5)	25(0/5)	2(0/5)	0(0/5)

Table 7: Example 4.1.1. Median false positives with median false negatives (strong/weak effects) in parentheses when the model is correctly specified.

	AIC	BIC	GAIC	GBIC	GBIC _p -L	GBIC _p
200	73(0/0)	0(0/4)	0(0/4)	0(0/4)	0(0/4)	0(0/4)
400	77(0/0)	0(0/4)	0(0/4)	0(0/4)	0(0/4)	0(0/5)
1600	4(0/2)	0(0/5)	0(0/5)	0(0/5)	0(0/5)	0(0/5)
3200	0(0/5)	0(0/5)	0(0/5)	0(0/5)	0(0/5)	0(0/5)

Table 8: Example 4.1.2. Median false positives with median false negatives in parentheses.

p	AIC	BIC	GAIC	GBIC	GBIC _p -L	GBIC _p
200	4(0)	4(0)	4(0)	3(0)	0(0)	0(0)
400	5(0)	5(0)	5(0)	5(0)	4(0)	0(0)
1600	8(0)	8(0)	8(0)	8(0)	7(0)	0(0)
3200	8(0)	8(0)	8(0)	8(0)	8(0)	0(0)

Table 9: Example 4.1.3. Median false positives with median false negatives in parentheses.

p	AIC	BIC	GAIC	GBIC	GBIC _p -L	GBIC _p
200	29(0)	1(0)	11(0)	1(0)	0(0)	0(0)
400	25(0)	5(0)	14(0)	1(0)	0(0)	0(0)
1600	19(0)	18(0)	14(0)	3(0)	1(0)	0(0)
3200	18(0)	17(0)	13(0)	9(0)	1(0)	0(0)